Mohamed El Merouani

Statistique Descriptive

(Cours et Exercices)

Première Edition - 2009

Mohamed El Merouani

Enseignant Chercheur
Département de Statistique et Informatique
Faculté Polydisciplinaire
Tétouan

Statistique Descriptive
(Cours et exercices)
1ère année universitaire

Première édition, 2009

Titre : Statistique Descriptive (Cours et exercices)

Auteur : Mohamed El Merouani

Année : 2009

Dépôt légal : 2009 MO 3071

Imprimerie et Papeterie : Al Khalij Al Arabi - Tétouan

Tél : 05 39 71 02 25 / Fax : 05 39 71 05 37

Sommaire

ETUSUP
.com

ETUSUP
.com

Avant-Propos

Cet ouvrage de Statistique Descriptive s'adresse aux étudiants de la première année universitaire, premier et deuxième semestres de toutes les filières de la nouvelle génération des Licences Fondamentales en sciences, sciences économiques et gestion...ainsi qu'aux étudiants de la première année des écoles d'ingénieurs et des écoles supérieures de commerce et gestion.

C'est un support de cours et des exercices, présentant les notions de base de la Statistique. Il dispense, donc, les étudiants de prendre des notes de cours, néanmoins, il les invite à faire un bon nombre d'exercices, extraits des contrôles et des examens, présentés à la fin de chaque chapitre.

M. El Merouani

Chapitre 0 :

Introduction à la Statistique

I.-Origine et définitions:

Le terme **STATISTIQUE** provient du mot latin **STATUS** due aux premières activités relatives à **l'Etat** telles que les recensements de la population, la collecte et la quantification de l'information,... etc. De-là provient l'utilisation du terme STATISTIQUE (ou STATISTIQUES au pluriel) pour se référer à une quantité d'information numérique collectée auprès de plusieurs situations telles que les accidents de la circulation, les indices de la bourse, les indices des prix, le Chômage, les votes aux élections,... etc. que nous sommes habitués de lire ou d'entendre dans les moyens de communication.

Nous donnons, ci-dessous, deux manières de définir la Statistique ou les Sciences Statistiques que nous considérons les plus générales et en même temps qui incluent les composantes basiques de la plupart des théories Statistiques :

V. Barnett (1973), « Comparative Statistical Inference », John Wiley & Sons : ((La Statistique se définie comme l'étude et la

recherche de l'information qui doit être utilisée dans l'orientation pour prendre une action dans une situation pratique qui comporte une incertitude)).

M. De Groot (1970), « Optimal Statistical Decision », McGraw Hill : ((Les Sciences Statistiques s'occupent du développement des théories et techniques appropriés pour faire des inférences sous les conditions d'incertitude et d'ignorance partielle qui existent nécessairement dans un large champ d'activités)).

Les deux définitions antérieures comprennent plusieurs termes que nécessitent une description et un commentaire plus détaillés, mais, nous, dans ce cours d'initiation, nous les surpassons pour développer l'aspect pratique de la Statistique.

II.-Termes usuels de la statistique :

1.-Population :

L'ensemble étudié par la statistique s'appelle Population. C'est l'ensemble statistique de référence sur lequel seront effectuées des applications quantitatives. On la note Ω.

Exemples :

- Ensemble des entreprises d'un secteur,
- Pièces produites par une usine,
- Ventes réalisées par un réseau de détaillants, etc.

2.- Individu:

C'est la partie élémentaire d'une population (définie selon les besoins de l'étude). On le note ω.

Exemples :

- Une entreprise,
- Une pièce produite,
- La vente d'un détaillant.

3.- Caractère:

C'est la caractéristique qui sert à décrire un individu d'une population. C'est donc une propriété commune des individus selon laquelle on va mener l'étude statistique. On le note X . Le caractère peut être quantitatif, mesurable ou qualitatif non mesurable. Dans le cas quantitatif, le caractère X s'appelle une variable statistique ou il est mesurable par une variable statistique. Alors, cette variable X peut être continue ($X \in IR$ ou à des intervalles de IR) ou discontinue c'est-à-dire discrète ($X \in IN$).

Exemples:

Chiffre d'affaires, valeur ajoutée, effectif des employés... sont des caractères quantitatifs appliqués à une entreprise.

Image de marque, climat social, solidité financière sont des exemples de caractères qualitatifs, non mesurables directement par une variable.

4.- Modalités :

Ce sont les différentes situations ou formes que peuvent prendre les individus selon le caractère étudié. On les note x_i . Les différentes modalités doivent être à la fois incompatibles, c'est-à-dire que chaque individu a une et une seule modalité, et exhaustives, c'est-à-dire les modalités englobent toute la population.

Exemples :

Les formes que peut prendre un caractère :

- bon ou mauvais pour une pièce contrôlée.
- les classes de chiffres d'affaires définies par le banquier pour estimer le niveau de risque.
- les classifications du personnel employé.
- etc.

5.- Effectifs ou fréquences absolues :

L'effectif est le nombre d'individus présentant chaque modalité ou le nombre de fois qu'une modalité est observée, on l'appelle aussi la fréquence absolue. On le note "n_i".

Exemple :

Les étudiants de la 1ère année éco. & gestion sont étudiés selon le nombre de leurs frères et sœurs. Alors on a :

- population : l'ensemble des étudiants de la 1ère année éco. & gestion.
- individus : étudiants.
- caractère : nombre des frères et sœurs (quantitatif et discret).
- modalités : 0, 1, 2, 3, 4, 5,…

- effectifs : 10, 20, 30, 40, 25, 12,…

6.- Effectif total :

L'effectif total est la somme des effectifs ; on le note N et on a :

$$N = n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i$$

(k étant le nombre des modalités).

7.- Fréquence relative:

La fréquence relative (ou fréquence tout court) est la proportion d'individus ayant la même modalité dans la population étudiée, on la note f_i. C'est le rapport entre l'effectif de la modalité et l'effectif total :

$$f_i = \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{n_i}{N}$$

(k étant le nombre des modalités).

Remarques:

- Lorsqu'on multiplie une fréquence relative par 100, on obtient le pourcentage de la modalité x_i dans le total de la population.
- La somme des fréquences relatives est égale à 1. En effet ;

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i}{N} = \frac{N}{N} = 1$$

Chapitre 1 :

Représentation des données

I.-Introduction :

Dans leur état brut, les informations ne peuvent être utilisées, d'où la nécessité de les "mettre en ordre", c'est-à-dire de les ranger.

A ce stade, elles forment une distribution ou une série statistique représentant un ensemble des modalités et des effectifs d'un caractère.

II.-Tableaux statistiques :

Le tableau correspondant aux séries à un seul caractère est à simple entrée, dans lequel figurent deux colonnes : l'une pour les modalités prise par la variables (x_i) et l'autre pour les effectifs correspondantes (n_i).

La présentation d'un tableau statistique fait correspondre à chaque modalité xi un "effectif" ou "fréquence absolue" ni.

1.- Cas des caractères qualitatifs :

Dans ce cas, les caractères ne sont pas mesurables, on peut donc les ranger dans un tableau statistique selon une logique qui permet au mieux l'interprétation.

Exemple :

Distribution des employés d'une multinationale selon leur nationalité :

Nationalité (x_i)	Effectifs (n_i)
Marocains	1450
Français	120
Tunisiens	50
Espagnoles	28
Autres	7
	N = 1655

2.- Cas des caractères quantitatifs :

Lorsque le caractère est quantitatif, il faut distinguer le cas des variables discrètes du cas des variables continues.

a.- Variables statistiques discrètes :

Dans ce cas, les valeurs xi sont discrètes et correspondent chacune à un effectif ni .

Exemple :

Soit la distribution statistique représentant le nombre des frères et sœurs des étudiants de la 1ère année éco. & gestion :

Frères et sœurs (x_i)	Effectifs (n_i)
0	10
1	20
2	30
3	40
4	25
5	12
6	10
7	8
8	4
9	2
10	1
	N = 162

b.- Variables statistiques continues:

Dans ce cas, les valeurs prises par la variable ne sont pas entières, ce qui nécessite la création de classes de valeurs possibles définies par les extrémités de classes (ou bornes).

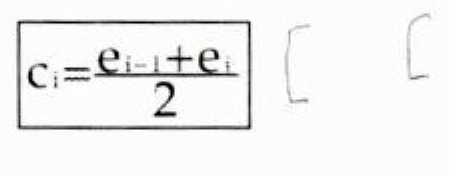
Exemple :

Soit la distribution statistique des employés d'une entreprise selon leurs salaires mensuels en dirhams :

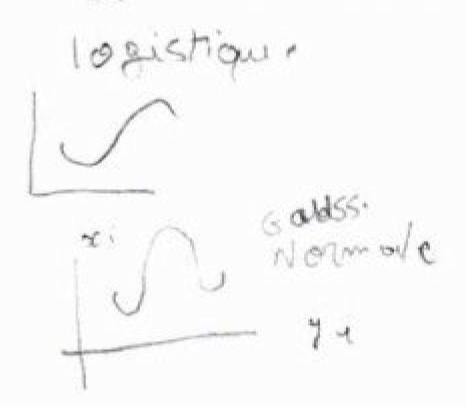
Salaires en DH (x_i)	Effectifs (n_i)
<2500	2
[2500, 3500[	4
[3500, 4500[	11
[4500, 5500[	5
>5500	3
	N = 25

Remarque :

- La division de la série en classes se fait selon la nature du sujet traité, elle relève donc du statisticien et de son appréciation du problème :
  - Si l'intervalle est petit, le nombre de classes sera important et peut rendre les calculs plus compliqués.
  - Si l'intervalle est grand, l'information se perd à cause de la condensation des chiffres.
- Dans le cas des variables statistiques continues, les modalités xi appartiennent à des intervalles de la forme $[e_{i-1}, e_i[$ que l'on appelle des classes et qui ont un centre :



$$c_i = \frac{e_{i-1} + e_i}{2}$$



- Ces intervalles ou classes ont une largeur ou amplitude :

$$a_i = e_i - e_{i-1}$$

Exemple :

Dans l'exemple précédent, les centres et amplitudes de classes sont :

$[e_{i-1}, e_i[$	n_i	ci	ai	fi	%
<2500	2	_	_	0,08	8%
[2500,3500[	4	3000	1000	0,16	16%
[3500,4500[	11	4000	1000	0,44	44%
[4500,5500[	5	5000	1000	0,2	20%
>5500	3	_	_	0,12	12%
	N = 25			$\sum_i f_i = 1$	100%

- On a complété le tableau, en y ajoutant les fréquences relatives, calculées comme on a vu, en utilisant la formule :

$$f_i = \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{n_i}{N}$$

- On a aussi calculé les pourcentages : à chaque fois, on a multiplié par 100 les fréquences relatives.

III.-Effectifs cumulés et fréquences cumulées :

1.- Effectifs cumulés croissants :

L'effectif cumulé croissant est le nombre de fois qui se présente une modalité ou plusieurs inférieures à une valeur donnée. On le note nic↑ ou nicc. Il se calcule en totalisant les effectifs avec ceux des classes suivantes. Il sert à répondre aux questions qui demandent un nombre ou un effectif « moins de » ou « inférieur à » une valeur donnée.

2.- Effectifs cumulés décroissants :

L'effectif cumulé décroissant est le nombre de fois qui se présente une modalité ou plusieurs supérieures à une valeur donnée. On le note nic↓ ou nicd. Il se calcule en totalisant les effectifs avec ceux des classes antérieures. Il sert à répondre aux questions qui demandent un nombre ou un effectif « plus de » ou « supérieur à » une valeur donnée.

3.- Fréquences cumulées croissantes :

Une fréquence cumulée croissante est la proportion des modalités « inférieur à » ou « moins de » une valeur donnée. On la note fic↑ ou ficc. Elle se calcule en totalisant les fréquences relatives avec celles des classes suivantes, ou en calculant le rapport de l'effectif cumulé croissant et l'effectif total :

$$fic\uparrow = \frac{nic\uparrow}{N}$$

4.- Fréquences cumulées décroissantes :

Une fréquence cumulée décroissante est la proportion des modalités « supérieur à » ou « plus de » une valeur donnée. On la note fic↓ ou ficd. Elle se calcule en totalisant les fréquences relatives avec celles des classes antérieures, ou en calculant le rapport de l'effectif cumulé décroissant et l'effectif total :

$$fic\downarrow = \frac{nic\downarrow}{N}$$

Exemple :

Pour l'exemple précédent, calculons les effectifs cumulés et les fréquences cumulées :

$[e_{i-1}, e_i[$	n_i	fi	nic↑	fic↑	nic↓	fic↓
<2500	2	0,08	2	0,08	25	1
[2500, 3500[	4	0,16	6	0,24	23	0,92
[3500, 4500[	11	0,44	17	0,68	19	0,76
[4500, 5500[	5	0,2	22	0,88	8	0, 32
>5500	3	0, 12	25	1	3	0, 12
	N = 25	$\sum_i f_i = 1$				

On peut donc dire d'après ces calculs que, par exemple, 68% des salariés gagnent moins de 4500 DH/mois.et 32% des employés ont un salaire supérieur à 4500 DH/mois.

Dans l'entreprise de cet exemple qui a un total de 25 employés, 17 parmi eux gagnent moins de 4500 DH/mois et 8 ont un salaire supérieur à 4500 DH/mois.

5.- Fonction de répartition :

Étant donnée une série statistique correspondante à la variable statistique X,

la fonction de répartition F(x) est définie par :

$$F(x) = f_i cc$$

avec $f_i cc$ est la fréquence cumulée croissante :

$$f_i cc = f_1 + f_2 + \cdots + f_i$$

et $x_{i-1} \leq x < x_i$ si X est discrète ou qualitative

ou $e_{i-1} \leq x < e_i$ si X est continue.

Donc $F(x) = \sum_{x<x_i} f_i$ si X est discrète ou qualitative

et $F(x) = \sum_{x<e_i} f_i$ si X est continue.

IV.-Représentations graphiques :

On a vu que l'on représente les séries statistiques par des tableaux, mais ces tableaux restent parfois insuffisants pour voir, par exemple, si une variable croit, décroît ou reste constante ; ou si elle présente autres caractéristiques. Les représentations graphiques peuvent nous donner une vue immédiate et complète des phénomènes étudiés.

Les représentations graphiques diffèrent selon la nature des variables étudiées.

1.- Caractères quantitatifs :

Dans ce cas, il faut faire la distinction entre les variables statistiques continues et discrètes.

a.- Variables statistiques discrètes :

***- Le diagramme en bâtons :**

Le diagramme en bâtons représente les fréquences f_i ou les effectifs n_i correspondants à chaque valeur de la variable x_i

Exemple :

Soit la série statistique donnant le nombre des frères et sœurs des 30 étudiants de la 1ère année éco. & gestion :

x_i	n_i	f_i
0	2	0,067
1	3	0,1
2	5	0,167
3	10	0,333
4	7	0,234
5	2	0,067
6	1	0,033
	N=30	1

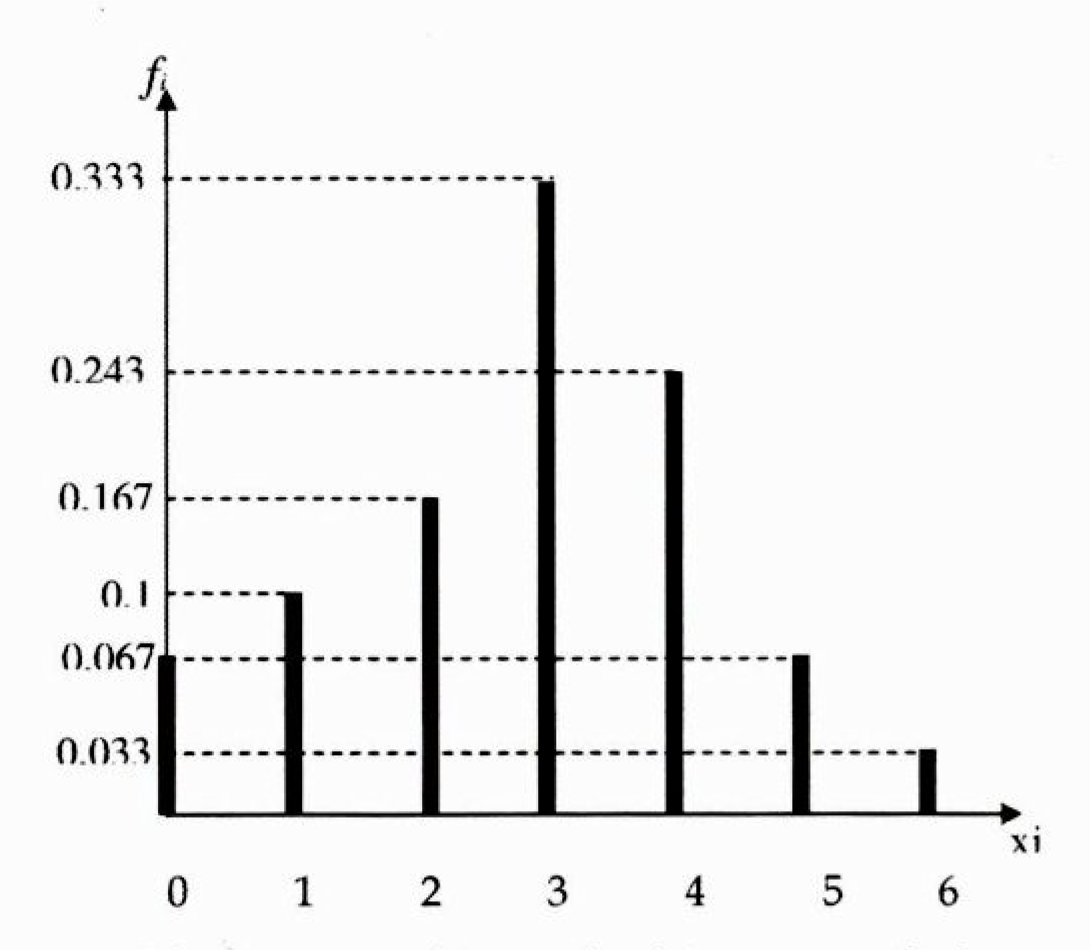


Diagramme en bâtons des fréquences relatives

***-La courbe cumulative :**

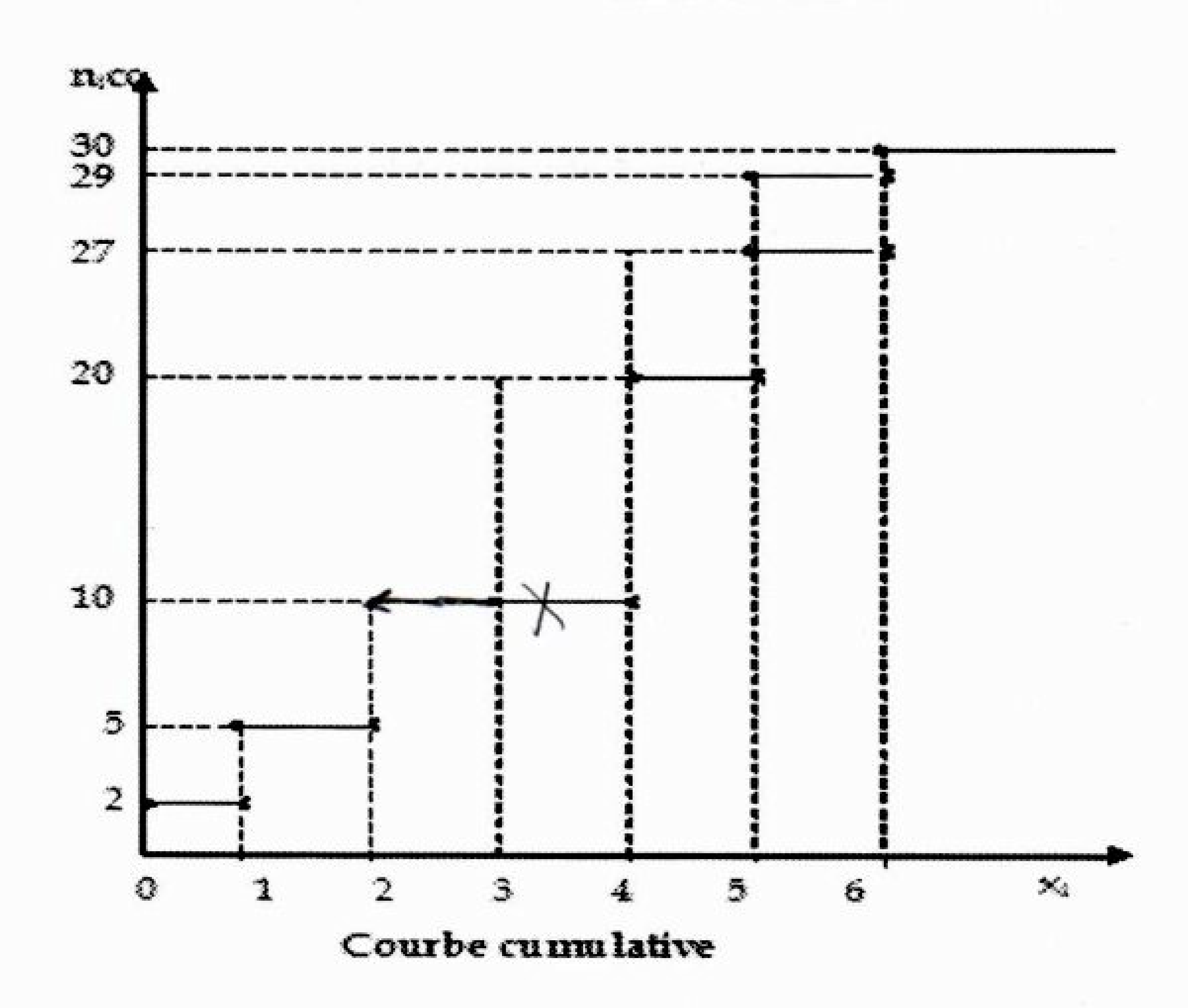
La courbe cumulative est une fonction de distribution qui représente les effectifs cumulés croissants $n_ic\uparrow$ en fonction des modalités x_i .

Elle permet de déterminer la proportion des individus de la population dont le caractère est inférieur à une certaine valeur x_i. Elle est représentée par une courbe en "escaliers" dont les paliers sont horizontaux.

Exemple :

Soit la série statistique quantitative discrète suivante (des nombres de frères et sœurs de 30 étudiants).

x_i	n_i	n_icc
0	2	2
1	3	5
2	5	10
3	10	20
4	7	27
5	2	29
6	1	30
	N=30	



Courbe cumulative

Pour la lecture du graphique, on dit par exemple que 20 étudiants ont moins de 4 frères et sœurs.

b.- Variables statistiques continues :

Comme pour ce type de variables, il y a une infinité de valeurs intermédiaires nécessitant le recours à des classes, on ne peut pas utiliser le diagramme en bâtons. On se propose d'utiliser la représentation des surfaces des classes par un histogramme.

***- L'histogramme :**

Pour tracer l'histogramme d'une variable statistique quantitative continue, on doit distinguer deux cas selon si les amplitudes des classes sont égales ou différentes.

***- Cas des amplitudes égales :**

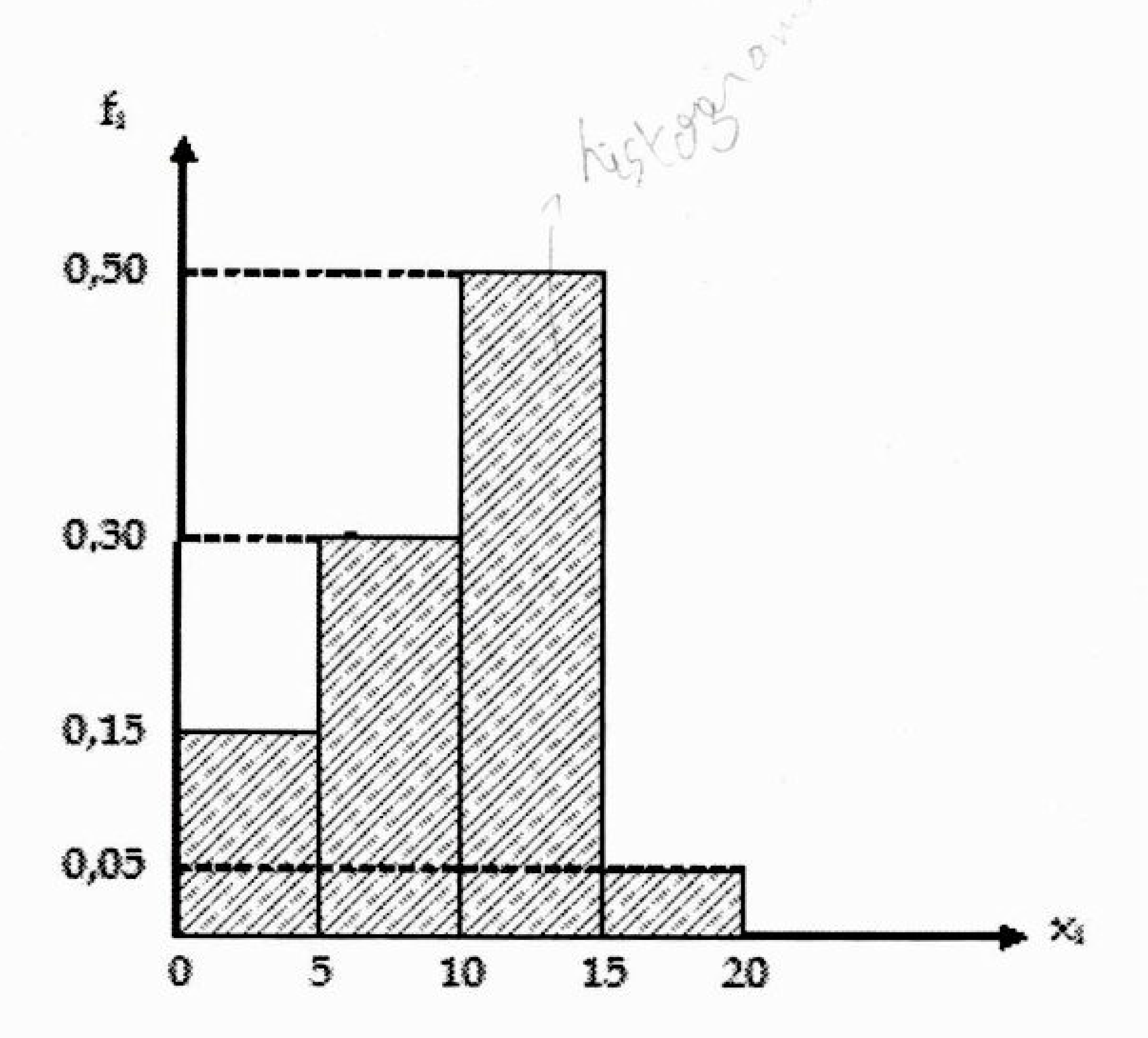
Exemple :

Soit la série représentant les notes des étudiants en Informatique :

$[e_{i-1}, e_i[$	n_i	f_i
[0,5[	60	0,15
[5,10[	120	0,30
[10,15[	200	0,50
[15,20[	20	0,05
	N=400	$\sum_i f_i = 1$

$$f_i = \frac{n_i}{N}$$

On remarque que l'amplitude de $a_i = e_i - e_{i-1} = 5$ pour toutes les classes.



Histogramme des fréquences relatives

L'histogramme est la surface hachurée.

On peut aussi représenter un histogramme pour les ni

<u>*- Cas des amplitudes différentes :</u>

<u>Exemple :</u>

Le salaire annuel des 100 employés d'une entreprise se répartit comme suit :

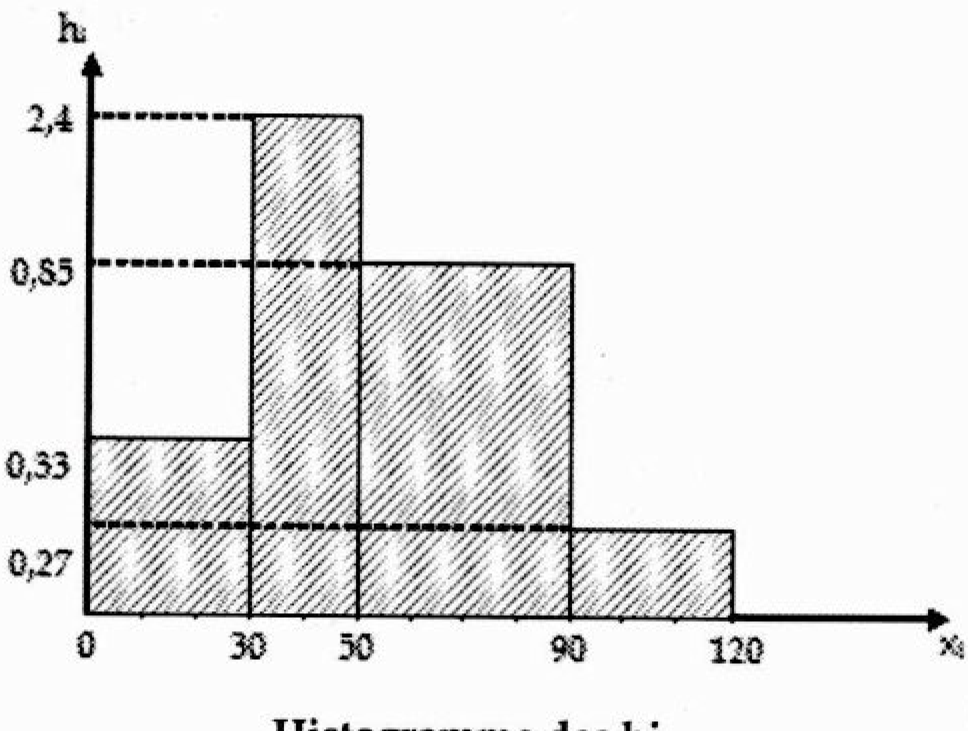
Salaire annuel en 1000 Dh	Nombre d'employés
[0,30[	10
[30,50[	48
[50,90[	34
[90,120[	8
Total	100

Représentons l'histogramme de cette distribution.

Pour homogénéiser la distribution et tenir compte de l'inégalité des amplitudes, on dresse le tableau suivant :

$[e_{i-1}, e_i[$	n_i	$f_i = n_i / N$	$a_i = e_i - e_{i-1}$	$h_i = n_i / a_i$
[0,30[	10	0,1	30	0,33
[30,50[	48	0,48	20	2,4
[50,90[	34	0,34	40	0,85
[90,120[	8	0,08	30	0,27
	N=100	1		

Alors, on représente l'histogramme pour h_i (hauteur des rectangles).



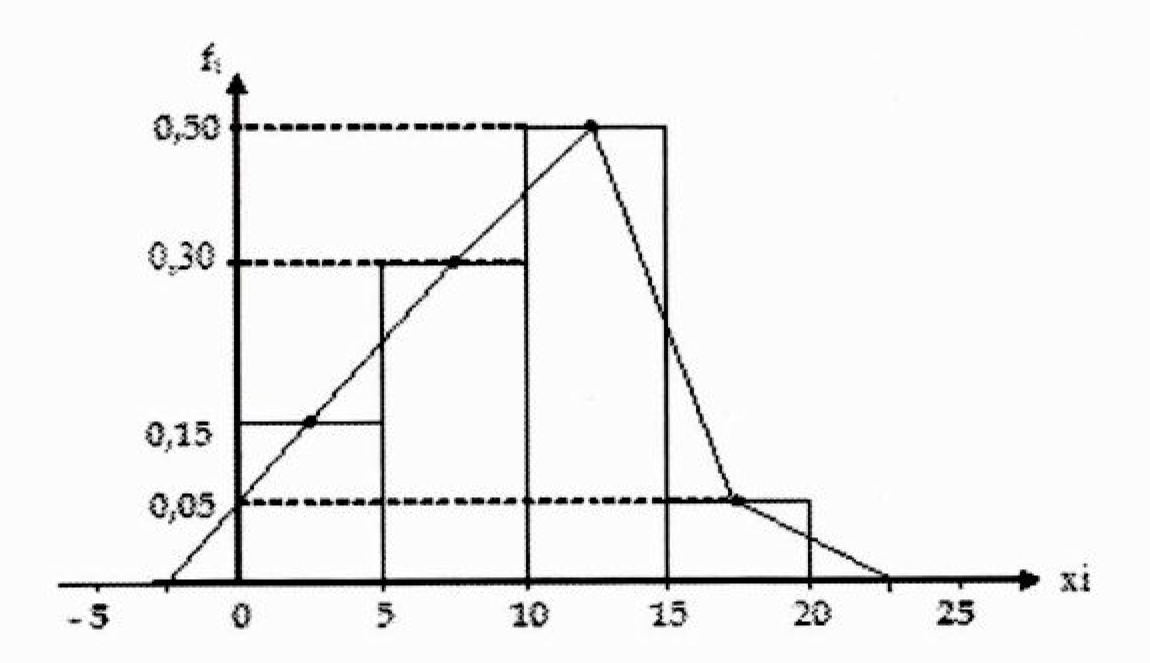
Histogramme des hi

***- Polygones des fréquences :**

Partant de l'histogramme et joignant par des segments de droite les milieux des sommets des rectangles, on obtient le polygone (des effectifs ou des fréquences).

Exemple :

Reprenons l'exemple des notes des étudiants en Informatique I, lorsque les classes ont des amplitudes égales :



Polygones des fréquences

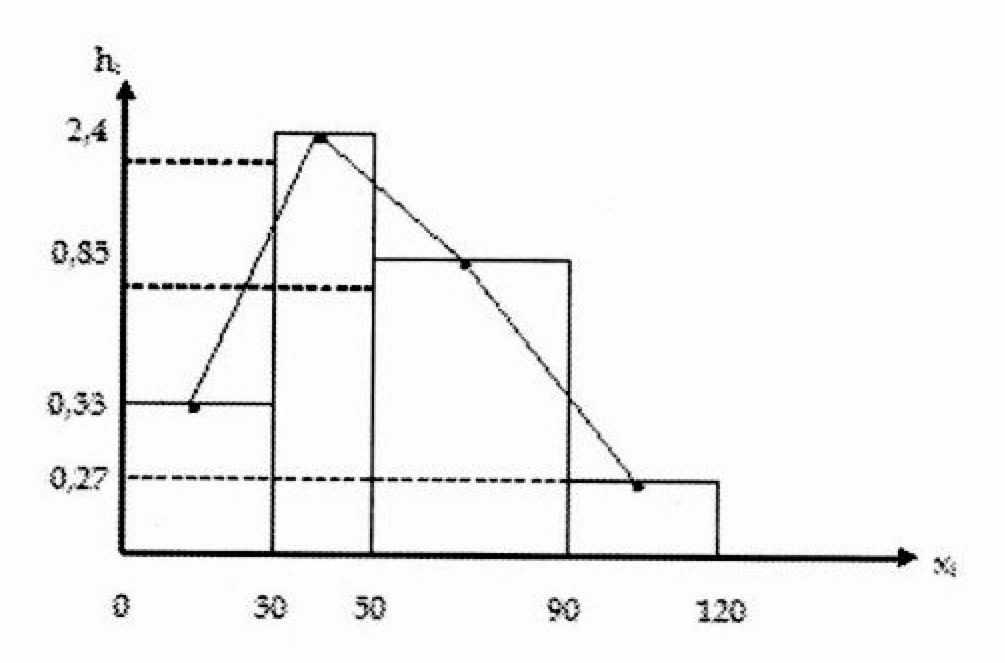
Il est préférable d'ajouter deux classes fictives aux extrémités de la série pour obtenir un polygone parfait

Remarque :

Il y a toujours conservation de la surface puisque l'aire sous le polygone est égale à 1 (en fréquence) (voir parties hachurées sur le graphique qui se compensent).

Exemples :

Pour l'exemple des salaires annuels des 100 employés d'une entreprise on a :



Remarque :

En faisant un ajustement graphique du polygone des fréquences, ce dernier tend vers une courbe continue appelée "courbe des fréquences".

2.- Les caractères qualitatifs :

a.- Les graphiques en tuyaux d'orgue :

Ce type de graphique représente en abscisse les différents caractères (x_i), de base constante, et en ordonnées la hauteur correspondant aux effectifs ou aux fréquences.

Généralement, on a tendance à ordonner les (x_i) selon un ordre décroissant en partant de l'origine des axes.

Exemple :

Les participants aux jeux olympiques selon le continent d'origine :

x_i	n_i
Afrique	80
Amérique	100
Asie	140
Australie	30
Europe	130

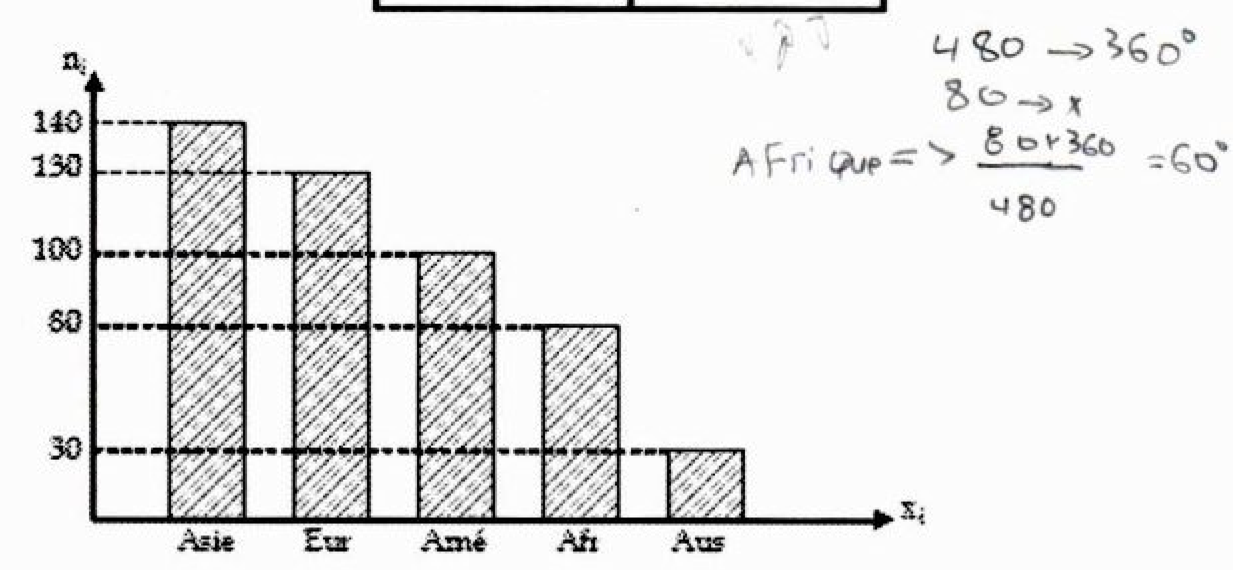


Diagramme en tuyaux d'orgue

***- Diagramme à secteurs :**

C'est un diagramme visualisant les parts relatives dans des secteurs de cercles. Chaque secteur correspond à une modalité, l'angle au centre est égal au produit de 360° par la fréquence f_i.

Exemple :

Reprenons l'exemple précèdent.

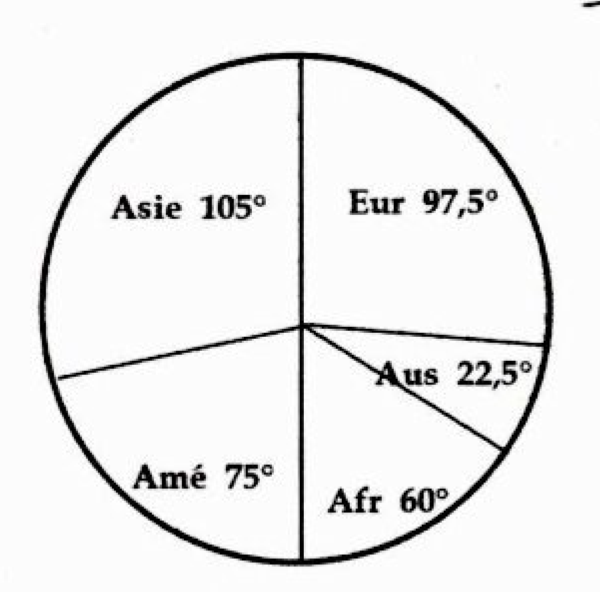
Le nombre total des participants est de 480.

- L'Afrique correspond à $\frac{80}{480} \times 360 = 60°$

$$\left[\text{Règle de trois} \quad \left.\begin{array}{l} 480 \to 360^\circ \\ 80 \to x \end{array}\right\} \Rightarrow x = \frac{80 \times 360}{480} = 60 \right]$$

L'Amérique $\rightarrow \dfrac{100}{480} \times 360 = 75^\circ$

- L'Asie $\rightarrow \dfrac{140}{480} \times 360 = 105^\circ$
- L'Australie $\rightarrow \dfrac{30}{480} \times 360 = 22{,}5^\circ$
- L'Europe $\rightarrow \dfrac{130}{480} \times 360 = 97{,}5^\circ$



Exercices

Exercice n°1 :

Pour chacune des distributions statistiques S1 et S2 suivantes:

1) Définir la population statistique, l'unité statistique, le caractère(en précisant sa nature (quantitatif ou qualitatif), la variable statistique et sa nature (si elle existe), le nombre de modalités que présente le caractère.

2) Construire (pour les distributions à caractère quantitatif) le diagramme en bâton ou histogramme et courbes cumulatives.

3) Construire pour les distributions à caractère qualitatif les diagrammes à secteurs.

Tableau S1 :

Répartition des hôtels homologués par le commissariat Général au tourisme au 1er janvier 1968 selon leur catégorie :

Catégorie (nombre d'étoiles)	Nombre d'hôtels
1 étoile	8346
2 étoiles	3770
3 étoiles	1161
4 étoiles	259
5 étoiles	48

qualitative

Tableau S2 :

Sur un rapport de 50 pages, on a relevé le nombre de fautes de frappe. On a obtenu le tableau suivant :

Nombre de fautes de frappe	Nombre de pages
1	8
2	12
3	15
4	8
5	4
6	3

Exercice n°2 :

Les salaires mensuels en dirhams des 50 employés d'une entreprise sont donnés par le tableau suivant :

Salaires en DH	Nombre d'employés
0 – 1500	3
1500 – 3000	15
3000 – 4500	20
4500 – 6000	8
6000 – 7500	4

1) Calculer :
   a) Tous les types de fréquences.
   b) Les centres des classes.
2) Représenter l'histogramme et le polygone des fréquences.
3) Quel est le pourcentage des employés qui ont un salaire :

a) Inférieur à 3000 DH/mois ?

b) Supérieur à 4500 DH/mois ?

c) Entre 3000 et 4500 DH/mois ?

d) Inférieur à 2800 DH/mois ?

Exercice n° 3 :

La distribution des salaires mensuels dans une entreprise est donnée par le tableau suivant (en Dhs) :

$[e_{i-1} ; e_i[$	n_i
<5000	12
[5000 ; 6000[	15
[6000 ; 7000[	27
[7000 ; 8000[	33
[8000 ; 9000[	30
[9000 ; 10000[	20
>10000	13

1) Calculer les fréquences relatives et les fréquences relatives cumulées croissantes et décroissantes.

2) Quel est le pourcentage des salariés qui gagnent moins de 7000 Dhs/mois ?

3) Quel est le pourcentage des salariés qui gagnent plus de 8000 Dhs/mois ?

Chapitre 2 :

Caractéristiques de tendance centrale

I.-Introduction :

Les caractéristiques des distributions statistiques est un outil essentiel de comparaison, se réfèrent à la tendance centrale (position), à la dispersion et à la forme (concentration).
Les caractéristiques de position ou de tendance centrale sont le mode, la médiane et les moyennes.

II.- Le mode (Mo) :

Le mode est la valeur ($M_o=x_i$) la plus fréquente de la variable, c'est-à-dire celle pour laquelle :

- La fréquence est maximale, ou
- L'effectif est le plus grand.

1.- Cas de la variable quantitative discrète :

Le mode correspond à la valeur x_i de la variable ayant la fréquence la plus élevée dans un tableau statistique.

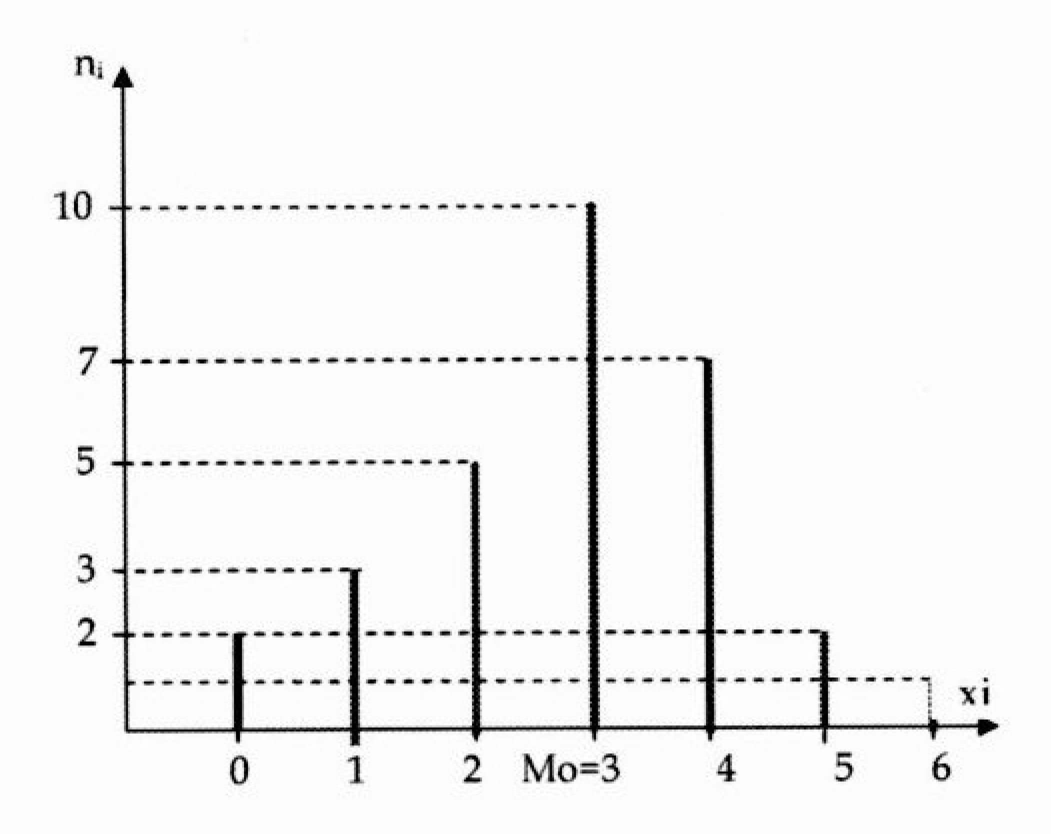
Exemple :

Soit la distribution statistique représentant le nombre de frères et sœurs des étudiants,

x_i	n_i
0	2
1	3
2	5
3	10 ←
4	7
5	2
6	1
	N = 30

$M_0=3$ puisque l'effectif le plus grand est égal à 10.

Graphiquement, le mode correspond à la valeur x_i possédant le plus haut bâton dans un diagramme en bâtons.



2.- Cas de la variable quantitative continue :

Dans ce cas, les données sont groupées en classes et le mode est représenté par la classe modale correspondante au plus grand effectif. Mais, il faut distinguer deux sous cas, lorsque les amplitudes des classes sont égales et lorsque les amplitudes sont différentes.

Exemple pour le cas des amplitudes égales :

Soit la série représentant les notes des étudiants :

Toutes les a_i égale à 5.

$[e_{i-1}, e_i]$	n_i
[0, 5[	60
[5, 10[	120

[10, 15[	200 ←
[15, 20[	20

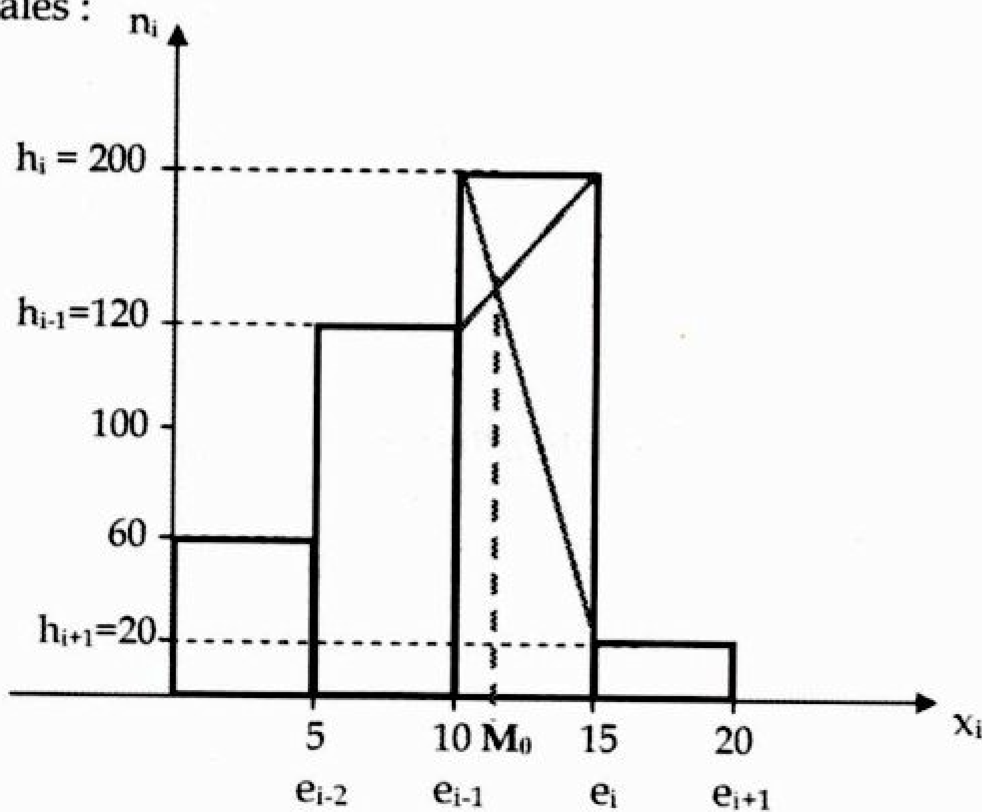
La classe modale est [10,15[, puisqu'elle correspond à l'effectif le plus grand.

le mode correspond :

- au centre de la classe modale :

$$M_0 = \frac{10+15}{2} = \frac{25}{2} = 12{,}5$$

- ou graphiquement, à une valeur donnée par la méthode des diagonales :



et on a la formule :

$$M_0 = e_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i \qquad (I)$$

ou la formule :

$$M_0 = e_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i+1}) + (h_i - h_{i-1})} a_i \qquad (II)$$

par application de (I), on a :

$$M_0 = 10 + \frac{20}{120+20}5 = 10 + \frac{100}{140} = 10 + 0{,}714 \Rightarrow \boxed{M_0 = 10{,}714}$$

Exemple pour le cas des amplitudes différentes :

Lorsque les amplitudes de classes sont différentes, on définit la classe modale comme étant celle du plus grande hauteur dans l'histogramme ($h_i = \frac{n_i}{a_i}$), (h_i représente la fréquence moyenne par unité d'amplitude). Il se peut avoir plusieurs classes modales. Dans la classe modale, il faut choisir ou sélectionner un point comme mode. Ce point se trouve donc dans la zone de la classe où il y a le plus d'observations.

Soit la série statistique :

Fréquence Moyenne par Unité d'amplitude

$[e_{i-1}, e_i[$	n_i	a_i	$h_i=n_i/a_i$
[100, 200[	15	100	0,15
[200, 400[	25	200	0,125
[400, 600[	40	200	0,2
[600, 700[	35	100	0,35 ←
[700, 1000[	20	300	0,067

La classe modale est [600, 700[(et non pas [400, 600[qui a n_i=40 qui est le plus grand) car [600, 700[correspond à h_i le plus élevée.

On applique la formule :

$$M_0 = e_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i$$

on a : $e_{i-1}=600$; $e_i=700$; $a_i=100$

$h_{i-1}=0,2$ et $h_{i+1}=0,067$

donc $$M_0 = 600 + \frac{0,067}{0,2+0,067} 100 \approx 625,1$$

Remarque:

Le mode ne doit être retenu que s'il est unique (série unimodale). Lorsque la distribution est multimodale, le mode perd toute signification.

III.- La médiane (Mé) :

1.- Définition :

La médiane est la valeur de la variable statistique qui partage la série statistique en deux parties de total des effectifs égaux, en supposant au préalable que la distribution soit classée par ordre de valeurs croissantes ou décroissantes de la variable.

Deux cas se présentent :

1er cas : si le nombre d'observations est impair (2k+1), alors il suffit de déterminer la médiane qui sera le $(k+1)^{ème}$ terme.

Exemple :

Soient les notes obtenues par 7 étudiants à l'examen :

10, 12, 9, 13, 8, 7, 11

Classons cette série dans l'ordre croissant :

$$\underbrace{7, 8, 9}, \underset{\text{Mé}}{10}, \underbrace{11, 12, 13} \Rightarrow \text{Mé} = 10$$

Puisqu'on a sept termes, la médiane correspond à 10, car il y a trois valeurs à gauche de 10 et trois valeurs à sa droite.

2ème cas : si le nombre d'observations est pair 2k, alors la médiane est donnée par l'intervalle médian de borne inférieure la valeur $k^{ème}$ et de borne supérieure la valeur $(k+1)^{ème}$ de la série. On prend comme médiane la moyenne (la centre de l'intervalle) des bornes de l'intervalle médian.

Exemple :

Soit maintenant la série suivante :

$$\overbrace{5, 6, 7, 8}, 9, 10, \overbrace{11, 12, 13, 14}$$

il y a dix valeurs, $10 = 2 \times 5 \Rightarrow k = 5^{ème}$ et $k+1=6^{ème}$, le $5^{ème}$ est 9 et le $6^{ème}$ est 10

$$\Rightarrow Mé = \frac{9+10}{2} = 9{,}5 \notin IN$$

Comme on est dans le cas discret et $9{,}5 \notin IN$, on ne peut pas la prendre comme une vraie médiane.

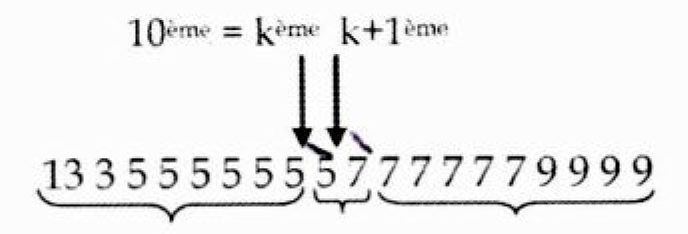
2- Méthode pratique de recherche de la médiane :

Exemple 1 : Soit la série statistique suivante :

x_i	n_i	n_icc
1	1	1
3	2	3
5	7	10 *
7	6	16
9	4	20
	N=20	

$\frac{N}{2} = \frac{20}{2} = 10$, valeur existante parmi les n_icc

$$\Rightarrow Mé = \frac{5+7}{2} = 6$$



Exemple 2 : Soit la série statistique suivante :

x_i	n_i	n_icc
1	1	1
3	3	4
5	7	11 *
7	3	14
9	6	20
	N=20	

$\frac{N}{2} = \frac{20}{2} = 10$, valeur existante parmi les n_icc

$\Rightarrow$ Mé = 5

10ème et 11ème

1 3 3 3 5 5 5 5 5 5 5 7 7 7 9 9 9 9 9 9

$$\Rightarrow \text{Mé} = \frac{5+5}{2} = 5$$

Lorsque la série statistique n'est pas unitaire, c'est-à-dire lorsque les effectifs sont différents de 1 ($n_i \neq 1$), alors si $N = \sum_i n_i$ est l'effectif total, on caractérise la médiane comme étant la valeur qui a la moitié des observations (N/2) inférieures à elle.

On cherche la valeur N/2 parmi les effectifs cumulés croissants, deux cas peuvent se présenter, soit il existe un n_{ic} tel que $N/2=n_{ic}$, soit aucun n_{ic} ne vérifie $N/2=n_{ic}$.

Dans le cas où il existe un $n_{ic}=N/2$, on sera dans la même situation que celle du cas d'un nombre pair d'observations, alors les deux positions centrales de la série ordonnée sont x_i et x_{i+1}, on prend, comme médiane la valeur moyenne $\text{Mé} = \frac{x_i + x_{i+1}}{2}$.

Dans le cas où $n_{ic} \neq N/2$ pour tout les n_{ic}, alors la valeur x_i correspondante à l'effectif cumulé où pour la 1ère fois $n_{ic} > N/2$, occupe l'unique position centrale ou les deux positions centrales.

On prend alors comme médiane Mé = x_i ou Mé = $\frac{x_i + x_i}{2}$ c'est-à-dire que Mé = xi.

quanti discr

Exemple 3 : Soit la série statistique suivante :

$$5, 6, 6, 8, 9, \underset{\|}{\overset{\text{Mé}}{}}10, 10, 10, 11, 12, 13$$

x_i	n_i	n_icc
5	1	1
6	2	3
8	1	4
9	1	5
10	3	8*
11	1	9
12	1	10
13	1	11
	N=11	

$N=11 \Rightarrow N/2= 5,5$ n'existe pas parmi les n_icc

il est dépassé pour la 1ère fois en 8
$\Rightarrow$ 10 = Mé

3.- Cas d'une variable continue :

Dans ce cas, les valeurs sont regroupées en classes. La médiane se calcule d'abord par détermination de la classe médiane $[e_{i-1}, e_i]$ qui correspond à la valeur de $n_{ic} = N/2$ ou à la 1ère valeur des n_{ic} qui dépasse N/2 dans le cas ou tous les n_{ic} sont différents de N/2.

Exemple 1 :

Soit la série statistique suivante :

[ei-1, ei[	n_i	n_icc
[0, 10[	48	48
[10, 20[	52	100 *
[20, 30[	60	160
[30, 40[	40	200
	N=200	

N/2 = 200/2 =100, cette valeur existe exactement parmi les n_{ic}

⇒ la classe médiane (c'est-à-dire l'intervalle qui contient la médiane) est [10, 20]=[e_{i-1}, e_i] et on prend Mé= e_i =20.

Exemple 2 :

Les salaires annuels des employés d'une entreprise (en milliers de DH) sont distribués comme suit:

[ei-1, ei[	n_i	n_icc
[40, 50[	12	12
[50, 60[	14	26
[60, 70[	20	46
[70, 80[	30	76 *
[80, 90[	14	90
[90,100[	10	100
	N=100	

N/2 = 100/2 =50, cette valeur n'existe pas parmi les n_icc

⇒ la classe médiane est la classe qui correspond à n_{ic} qui dépasse le premier la valeur 50 c'est-à-dire [70, 80]=[e_{i-1}, e_i] et ensuite, on applique la formule :

$$\text{Mé} = e_{i-1} + \frac{\frac{N}{2} - n_{(i-1)cc}}{n_i} a_i$$

$$\text{Mé} = 70 + \frac{50-46}{30} 10 = 71{,}33 \in [70{,}80]$$

Interprétation :

On a : N/2=50 et Mé = 71,33 . 10^3 DH = 71330 DH

Il y a 50 employés qui ont un salaire inférieur à 71330 DH et 50 autres qui ont un salaire supérieur à 71330 DH.

4.- Méthode graphique de détermination de la médiane :

On trace la courbe des fréquences (ou des effectifs) cumulées croissantes et la courbe des fréquences (ou des effectifs) cumulées décroissantes sur le même repère. La projection du point d'intersection de ces deux courbes sur l'axe des abscisses donne la valeur de la médiane.

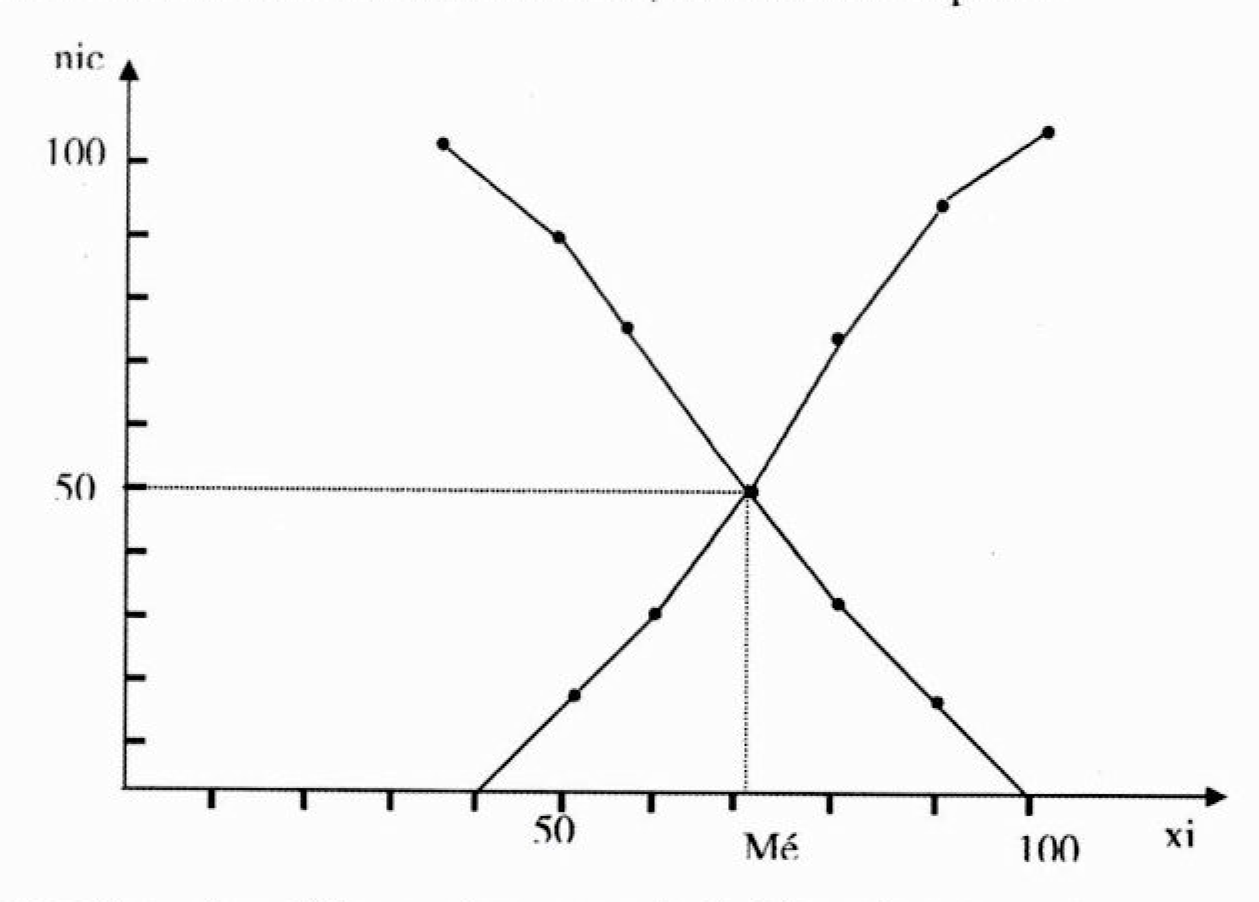
Exemple :

Soit la distribution de l'exemple précédent :

[e_{i-1}, e_i[	n_i	n_ic↑	n_ic↓
[40, 50[	12	12	100
[50, 60[	14	26	88
[60, 70[	20	46	74
[70, 80[	30	76	54

[80, 90[	14	90	24
[90,100[	10	100	10
	N=100		

Traçons la courbe des $n_ic\uparrow$ et des $n_ic\downarrow$ sur le même repère :



on voit que la médiane est presque égale à la valeur trouvée algébriquement **71,33**

IV.- Les moyennes :

Généralement, il y a quatre types de moyennes : arithmétique, géométrique, harmonique et quadratique.

1.- Moyenne arithmétique ($\overline{x}$) :

La moyenne arithmétique d'une variable statistique est égale à la somme des valeurs de la variable pondérées par les fréquences relatives.

$$\overline{x} = \sum_{i=1}^{k} f_i x_i$$

donc $$\overline{x} = \sum_{i=1}^{k} \frac{n_i}{N} x_i = \frac{1}{N} \sum_{i=1}^{k} n_i x_i$$

d'où $$\overline{x} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i$$

Exemple :

Soient les notes d'un groupe de 30 étudiants

x_i	n_i	$n_i x_i$	f_i	$f_i x_i$
7	4	28	0,133	0,931
8	5	40	0,167	1,336
9	6	54	0,2	1,8
13	7	91	0,233	3,029
14	6	84	0,2	2,8
15	2	30	0,067	1,005
Total	N=30	327	$\sum_i f_i = 1$	10,901

La note moyenne de ce groupe est :

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i = \frac{1}{30} \cdot 327 = 10{,}9 \text{ sur } 20$$

$$\overline{x} = \sum_{i=1}^{k} f_i x_i = 10{,}901 \approx 10{,}9$$

***- Variable statistique continue :**

Pour le calcul de la moyenne arithmétique dans le cas de la variable statistique continue, il faut procéder par la formule suivante :

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{k} n_i c_i$$

où c_i représente le centre des classes.

Exemple :

L'âge des salariés d'une entreprise est distribué comme suit :

$[e_{i-1}, e_i[$	n_i	c_i	$n_i c_i$
[10, 20[	2	15	30
[20, 30[	15	25	375
[30,40[	33	35	1155
[40, 50[	13	45	585
plus de 50	7	55 ?	385
Total	N=70		2530

Lorsque la borne n'est pas définie, il faut la définir soit par le "bon sens" soit par extrapolation des données. Ici, on peut dire que :

- L'âge de la retraite est de 60 ans, donc la classe "plus de 50 ans" est : [50, 60[,
- On a des classes d'amplitude 10, donc la dernière sera de [50, 60[.

$$\text{Ainsi} \quad \overline{x} = \frac{1}{N}\sum_{i=1}^{k} n_i c_i = \frac{1}{70}\cdot 2530 \cong 36{,}14$$

36,14 est l'âge moyen des 70 salariés de l'entreprise.

***- Calcul de $\overline{x}$ par changement d'origine :**

Lorsque les valeurs xi et ni sont trop grandes et que le calcul devient volumineux, il est préférable de procéder à un changement d'origine par la transformation suivante :

Si c_{i_0} est une origine quelconque, soit le changement $x'_i = x_i - c_{i_0}$, alors la moyenne sera $\overline{x'} = \overline{x} - c_{i_0}$ d'où la moyenne arithmétique recherchée sera $\overline{x} = \overline{x'} + c_{i_0}$

Exemple :

Soit la série statistique suivante :

$[e_{i-1}, e_i[$	n_i	c_i	$c_i - c_3 = c'_i$	$n_i c'_i$
[400, 500[	8	450	-200	-1600
[500, 600[	10	550	-100	-1000
[600,700[	12	650	0	0
[700, 800[	50	750	100	5000
[800, 900[	20	850	200	4000
	N=100			6400

$c_{io} = c_3 = 650$; le choix de cette valeur s'est fait par le choix de la valeur centrale des observations.

$$\overline{x'} = \overline{x} - c_3 \Rightarrow \overline{x} = \overline{x'} + c_3$$

or $\overline{x'} = \frac{1}{100} \times 6400 = 64$ donc $\overline{x} = 64 + 650 = 714$

***- Propriétés de la moyenne arithmétique :**

1) La somme des écarts à la moyenne arithmétique est nulle : $\sum_{i=1}^{k} n_i (x_i - \overline{x}) = 0$

En effet :

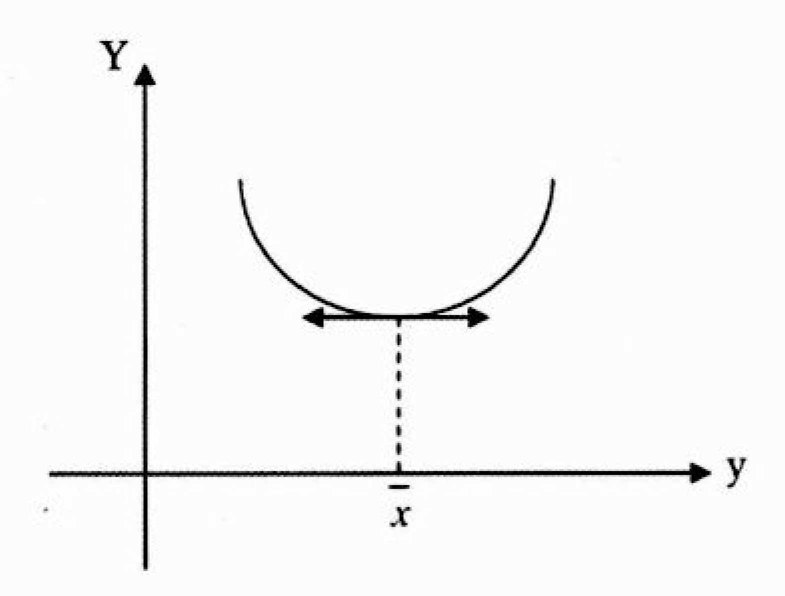
$$\sum_{i=1}^{k} n_i (x_i - \overline{x}) = \sum_{i=1}^{k} n_i x_i - \overline{x} \sum_{i=1}^{k} n_i = N\overline{x} - \overline{x}N = 0$$

2) La somme des carrées des écarts à $\overline{x}$ est minimale, c'est-à-dire : $\sum_{i=1}^{k} n_i (x_i - \overline{x})^2$ est minimale.

En effet :

soit la fonction $Y(y) = \sum_{i=1}^{k} n_i (x_i - y)^2$

on va démontrer que Y est de dérivée première nulle en $y_0 = \overline{x}$ et de dérivée seconde positive (concavité vers le haut)



on a : $Y(y) = \sum_{i=1}^{k} n_i (x_i^2 - 2x_i y + y^2)$

$$= \sum_{i=1}^{k} n_i x_i^2 - 2y\sum_{i=1}^{k} n_i x_i + y^2 \sum_{i=1}^{k} n_i$$

$$\frac{dY}{dy} = -2\sum_{i=1}^{k} n_i x_i + 2y\sum_{i=1}^{k} n_i = 0 \Leftrightarrow y\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i x_i$$

$$\Leftrightarrow y = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = \bar{x}$$

donc **Y** admet un extremum (optimum) en $\bar{x}$.

$$\frac{d^2Y(y)}{dy^2} = 2\sum_{i=1}^{k} n_i = 2N \succ 0$$

d'où Y admet un minimum en $\bar{x}$ (concavité vers le haut).

2- Moyenne géométrique G :

Elle est égale à la racine N^ième du produit des k valeurs d'une série statistique.

a- Cas d'une série simple :

Tous ses N valeurs sont d'effectifs égaux à 1, alors :

$$G = \sqrt[N]{x_1.x_2.........\ ...x_N} = (x_1.x_2............x_N)^{\frac{1}{N}}$$

$$G = \left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}$$

$$\log G = \frac{1}{N}\log\left(\prod_{i=1}^{N} x_i\right)$$

donc $$\log G = \frac{1}{N}\sum_{i=1}^{N}\log x_i$$

d'où on a supposé que $\forall$ i=1, …, N $x_i \geq 0$

b- Cas d'une série pondérée :

Ses valeurs $x_1,......, x_k$ ont des effectifs différents $n_1,......., n_k$ respectivement.

$$G = \sqrt[N]{x_1^{n_1}.x_2^{n_2}.........\ x_k^{n_k}} = \left(\prod_{i=1}^{k} x_i^{n_i}\right)^{\frac{1}{N}} = \prod_{i=1}^{k} x^{\frac{n_i}{N}}$$

$$N = n_1+n_2+............+n_k$$

$$G = \prod_{i=1}^{k} x_i^{f_i}$$

$$\Rightarrow \log G = \log\left(\prod_{i=1}^{K} x_i^{f_i}\right) = \sum_{i=1}^{k}\log x_i^{f_i} = \sum_{i=1}^{k} f_i \log x_i$$

ici, aussi, on a supposé que $x_i \geq 0$ $\forall$ i=1, …, k

Remarque :

La moyenne géométrique est utilisée pour le calcul des taux d'accroissements moyens, des moyennes de coefficients multiplicateurs…c'est-à-dire, dans les cas où la variable représente des variations cumulatives.

Exemple 1 :

Calculer la moyenne géométrique de la distribution de fréquence suivante :

x_i	n_i
1	2
2	5
3	3
	N = 10

On peut utiliser des logarithmes népériens ou décimaux sans changer le résultat de G.

Alors utilisons les logarithmes décimaux, on aura :

x_i	n_i	$\log x_i$	$n_i \log x_i$
1	2	0	0
2	5	0,301	1,505
3	3	0,4771	1,4313
Total	N=10		2,9363

Alors $$\log G = \frac{1}{N}\sum_{i=1}^{k} n_i \log x_i$$

$$= \frac{1}{10} \cdot 29363 = 0,29363$$

$$\Rightarrow \quad G = 10^{0,29363} = 1,97$$

Exemple 2 :

Le chiffre d'affaire d'un projet a produit les augmentations annuelles suivantes :

Année	Augmentation en %
1ère année	4%
2ème année	6%
3ème année	6%
4ème année	6%
5ème année	5%
6ème année	5%

L'augmentation moyenne annuelle est une moyenne géométrique

$$G = \sqrt[6]{(1,04)(1,06)^3(1,05)^2}$$
$$G = 1,05331$$

soit un taux de croissance de 5,331% approximativement.

En effet ; si C_o est le chiffre d'affaire au début de la 1ère année, à la fin de cette année il devient : $C_o + C_o \cdot \frac{4}{100} = C_o\,(1+0,04) = 1,04\,C_o = C_1$, ce chiffre C_1 devient à la fin de la 2ème année $C_1 + C_1 \cdot \frac{6}{100} = 1,06\,C_1 = C_2$ et ainsi de suite, à la fin de la 6ème année, le C_o devient

$C_F = 1,05 \,.\, 1,05 \,.\, 1,06 \,.\, 1,06 \,.\, 1,06 \,.\, 1,04\, C_o$

$$\Rightarrow \quad \frac{C_F}{C_O} = (1,05)^2 \cdot (1,06)^3 .1,04 = G^6$$

ainsi, l'augmentation moyenne annuelle est la moyenne éométrique

$$G = \sqrt[6]{(1,04)(1,06)^3(1,05)^2}$$
$$G = 1,05331$$

soit un taux de 5,331% .

3.- Généralisation de la moyenne :

L'expression de la moyenne peut se généraliser de plusieurs manières, une d'entre elles est ce que l'on appelle la moyenne d'ordre r, notée M_r, définie par :

$$(M_r)^r = \sum_{i=1}^{k} f_i c_i^r \Leftrightarrow M_r = \left(\sum_{i=1}^{k} f_i c_i^r\right)^{\frac{1}{r}}$$

Ainsi, la moyenne arithmétique est un cas particulier avec r =1, c'est-à-dire $M_1=\bar{x}$.

On peut montrer que $G = M_0$ par passage à la limite et par des théorèmes pratiques de dérivation des fonctions.

$$G = M_0 = \lim_{\varepsilon \to 0} M_\varepsilon = \lim_{\varepsilon \to 0}\left(\sum_{i=1}^{k} f_i c_i^\varepsilon\right)^{\frac{1}{\varepsilon}}$$

4.- Moyenne harmonique H :

On définit la moyenne harmonique comme la moyenne généralisée d'ordre -1. Elle se note H

$$H = M_{-1} = \left(\sum_{i=1}^{k} f_i x_i^{-1}\right)^{-1} = \left(\sum_{i=1}^{k} \frac{n_i}{N} x_i^{-1}\right)^{-1}$$

$$H = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^{-1}\right)^{-1} = N\left(\sum_{i=1}^{k} \frac{n_i}{x_i}\right)^{-1}$$

$$\boxed{H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}}$$

C'est la formule du cas d'une série pondérée.

***- Cas d'une série simple :**

$$\boxed{H = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i}}}$$

Remarques :

La moyenne harmonique se calcule pour des valeurs de la variable X non nulles et elle n'a de signification concrète que si l'inverse $\frac{1}{x_i}$ de la valeur x_i a un sens.

Elle est utilisée pour le calcul des moyennes de pourcentages, de ratios et de rapports, de même que pour l'étude du pouvoir d'achat (inverse du mouvement général des prix), etc.

Exemple :

Une imprimerie a imprimé 500 livres à la vitesse de 400 livres par heure et elle a imprimé 1000 livres identiques à la vitesse de 350 livres par heure. On veut déterminer la vitesse moyenne pour imprimer les 1500 livres.

Alors c'est $$H = \frac{1500}{\frac{n_1}{x_1} + \frac{n_2}{x_2}} = \frac{1500}{\frac{500}{400} + \frac{1000}{350}}$$

$$= \frac{1500}{1{,}25 + 2{,}86} \cong 365 \text{ livres/heure en}$$

moyenne.

5.- Moyenne quadratique Q :

La moyenne généralisée d'ordre r =2 s'appelle la moyenne quadratique.

$$Q = M_2 = \left(\sum_{i=1}^{k} f_i x_i^2\right)^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{k} \frac{n_i}{N} x_i^2}$$

$$\boxed{Q = \sqrt{\frac{1}{N} \sum_{i=1}^{k} n_i x_i^2}}$$

*- Cas d'une série simple :

$$Q = \sqrt{\frac{1}{N}\sum_{i=1}^{k} x_i^2}$$

Exemple :

Calculons Q de la série suivante :

x_i	n_i	x_i^2	$n_i\ x_i^2$
2	7	4	28
4	10	16	160
5	20	25	500
7	30	49	1470
9	15	81	1215
10	10	100	1000
11	8	121	968
	N=100		5341

$$Q=\sqrt{\frac{5341}{100}} = \sqrt{53{,}41} \cong 7{,}31$$

Remarque :

On peut montrer que les moyennes généralisées d'ordre r d'une même série statistique vérifient :

si $p \leq q$ alors $M_p \leq M_q$

ainsi $M_{-1} \leq M_0 \leq M_1 \leq M_2$

c'est-à-dire $H \leq G \leq \overline{x} \leq Q$

Exercices

Exercice n° 1 :

Au cours d'une enquête dans une entreprise, on a relevé les 20 salaires suivants: 1800, 2200, 2000 , 3800, 3500, 1700, 2600, 2800, 2200, 2600, 2400, 2300, 2100, 2400, 3500, 2600, 2400, 2100, 1800, 2600.

1) Calculer le salaire moyen $\bar{x}$.

2) Regrouper les 20 salaires en tableau, avec les classes ci-dessous:
[1500 ; 2000[; [2000; 2500[; [2500 ; 3000 [; [3000 ; 3500 [;
[3500 ; 4000 [

Tracer l'histogramme correspondant en précisant la hauteur de chaque classe. A quelle valeur $\bar{x}_1$ du salaire moyen conduit ce classement.

3) Si l'on regroupe les deux dernières classes préciser la déformation subie par l'histogramme précédent.

A quelle valeur $\bar{x}_2$ du salaire moyen conduit ce nouveau classement ? Conclusion?

Exercice n° 2 :

La répartition de 100 exploitations agricoles selon leur superficie, exprimée en hectares (ha), est donnée par le tableau suivant :

Superficies (en ha) comprises entre	Nombre d'exploitations
10 – 20	10
20 – 30	15
30 – 40	25
40 – 50	30
50 – 60	10
60 – 70	6
70 – 80	4

1) Calculer tous les types de fréquences.

2) Déterminer le mode et la médiane graphiquement et par le calcul. Donner leurs interprétations.
3) Quel est le pourcentage des exploitations agricoles qui ont une superficie :
   a) Inférieure à 30 ha ?
   b) Supérieure à 40 ha ?
   c) Comprise entre 30 et 40 ha ?
   d) Inférieure à 47 ha ?

Exercice n°3 :

Sur un rapport de 50 pages dactylographiées, on a relevé le nombre de fautes de frappe par page. On a obtenu le tableau suivant :

Nombre de fautes: xi	**Effectifs ni**
1	8
2	12
3	15
4	8
5	4
6	3

1) Que représente Σni
2) Calculer $\Sigma xini$ .En déduire le nombre moyen de fautes par page.

Exercice n°4 :

On se donne la série statistique suivante :1 , 2 , 5,7, 10, 13. En utilisant un coefficient de pondération identique calculer les moyennes arithmétique, géométrique, harmonique et quadratique.

Exercice n° 5 :

La distribution, en pourcentage, de 50 employés d'une entreprise selon leurs salaires mensuels (en dirhams) est donnée par le tableau suivant :

Salaires en DH compris entre	Pourcentages des employés
0 - 1500	6 %
1500 - 3000	30 %
3000 - 4500	40 %
4500 - 6000	16 %
6000 - 7500	8 %

1) Calculer les fréquences relatives et déduire les différents effectifs.
2) Calculer les fréquences relatives cumulées décroissantes et représenter leur courbe.
3) Quel est le salaire le plus fréquent ?
4) Quel est le salaire médian ?
5) Déterminer le salaire moyen par la méthode directe et par un changement d'origine convenable.

Exercice n° 6 :

Une entreprise produit un certain bien. Elle dispose de 3 usines A, B et C dont les rendements sont 20 unités par jour pour A, 32 pour B et 40 pour C.

L'usine A a produit 8000 unités, l'usine B a produit 9600 et C a produit 12000, pendant une période déterminée.

Quel est le rendement moyen de cette entreprise pendant cette période ? (Justifier le choix de la méthode utilisée).

Exercice n° 7 :

La durée de l'unité téléphonique de base entre 2 villes A et B est de 20 secondes; cette unité est facturée 50 centimes par l'administration.

Une durée inférieure à 20secondes génère 1 unité et coûte donc 50 centimes.

Une durée comprise entre [20s ; 40s [génère 2 unités et coûte 100 centimes et ainsi de suite.

Le tableau suivant représente la distribution de la durée X (en secondes) et du nombre d'unités Y correspondantes, des communications téléphoniques entre les 2 villes A et B.

Durée X (en seconde)	Fréquence fi (en %)	Nombre d'unités Yi
[120; 140[	2	7
[140 ; 160[	6	8
[160 ; 180[	12	9
[180 ; 200[	20	10
[200 ; 220[	18	11
[220 ; 240[	14	12
[240 ; 260[	10	13
[260 ; 280[	7	14

[280 ; 300[	5	15
[300 ; 320[	3	16
[320 ; 340[	2	17
[340 ; 360[	1	18

1) Quelle est la nature statistique de la variable **X**? Même question pour **Y.**

2) Déterminer le coût moyen d'une communication entre A et B .

3) Trouver la relation entre le centre **ci** de la i[ème] classe et la valeur **yi** de **Y** associée

En supposant que la distribution des durées est uniforme à l'intérieur des classes, calculer la durée moyenne d'une communication.

4) L'administration envisage de faire passer la durée de l'impulsion de base de 20s à 40 secondes. Une étude préalable a montré que ni le nombre ni la distribution de la durée des communications ne seraient affectés par une telle mesure.

Calculer la baisse des recettes engendrée par cette décision.

Chapitre 3 :

Caractéristiques de dispersion

I.- Introduction :

Les caractéristiques de tendance centrale vues précédemment ne nous permettent pas de faire la différence entre deux séries statistiques. En effet, deux séries peuvent avoir la même moyenne arithmétique et la même médiane sans qu'elles soient identiques. On introduit, alors, autres caractéristiques dites de dispersion qui estiment dans quelle mesure les observations s'écartent les unes des autres ou de leur valeur centrale.

Les caractéristiques de dispersion sont :

- l'Etendue,
- les Quantiles,
- l'Ecart absolu moyen,
- la variance et l'Ecart-type
- le coefficient de variation,

II.- L'Etendue :

1.- Définition :

L'Étendue ou l'Intervalle de Variation est la différence entre la valeur la plus faible et la valeur la plus élevée d'une série statistique.

2.- Exemple :

On considère les salaires des employés de deux entreprises A et B :

Pour A : 700, 720, 750, 800, 900, 1000, 1150

Pour B : 20, 100, 200, 800, 1300, 1600, 2000.

On a $\overline{X}_A = \overline{X}_B = 860$ et $Mé_A = Mé_B = 800$

Même si les moyennes arithmétiques et les médianes de ces deux séries sont identiques, on ne peut conclure qu'elles sont identiques car l'information sur le groupement/l'écart (la dispersion) des éléments des séries n'est pas disponible.

Calculons l'étendue pour ces deux entreprises :

Pour A, l'étendue est de 450=1150-700

Pour B, l'étendue est de 1980=2000-20

La dispersion est donc plus forte pour B que pour A.

3.- Exemple :

Sur 1000 employés d'une entreprise, l'étendue de l'âge est de 11 ans (49 ans - 38 ans), il suffit qu'un jeune employé (sur 1000) de 18 ans soit embouché pour que l'étendue passe à 31 ans (49 ans - 18 ans) ! On voit donc que l'étendue est une manière très simpliste de mesurer la dispersion.

III.- Les Quantiles :

1.- Définition :

Les quantiles sont les valeurs de la variable statistique qui partagent la distribution en "n" parties composées du même effectif "$\frac{N}{n}$".

Exemple :

La médiane est un quantile qui partage la distribution en "2" parties composées du même effectif "$\frac{N}{2}$".

Remarques :

- Les quantiles en eux-mêmes sont des caractéristiques de position (non centrale), tandis que l'intervalle mesurant l'écart entre deux quantiles et une caractéristique de dispersion.
- Pour la détermination des quantiles, on suppose que les valeurs de la série statistique sont classées dans un ordre croissant ou décroissant.

2.- Les quartiles :

a.- Définition :

Les quartiles sont les trois valeurs que l'on note Q_1, Q_2, Q_3 de la variable statistique qui partagent la distribution en "4" parties composées du même effectif "$\frac{N}{4}$".

← 25% →← 75% →

|————|————|————|————|

← $\frac{N}{4}$ →← $\frac{N}{4}$ →← $\frac{N}{4}$ →← $\frac{N}{4}$ →

← Q_1 → Q_2=Mé ← 50% →

50%

← 75% → Q_3 ← 25% →

Remarque :

Il y a $N\frac{i}{4}$ observations à gauche (inférieurs) de Q_i, i=1, 2, 3

b.- L'intervalle interquartile :

C'est la différence entre Q_3 et Q_1. Soit $Q_3 - Q_1$; il contient 50% des valeurs de la variable présentés par la moitié centrale des effectifs observés. Cet intervalle élimine l'influence des valeurs extrêmes.

c.-La dérivation quartile ou le semi-interquartile est égale à

$$\frac{Q_3 - Q_1}{2}$$

d.-L'écart interquartile relatif :

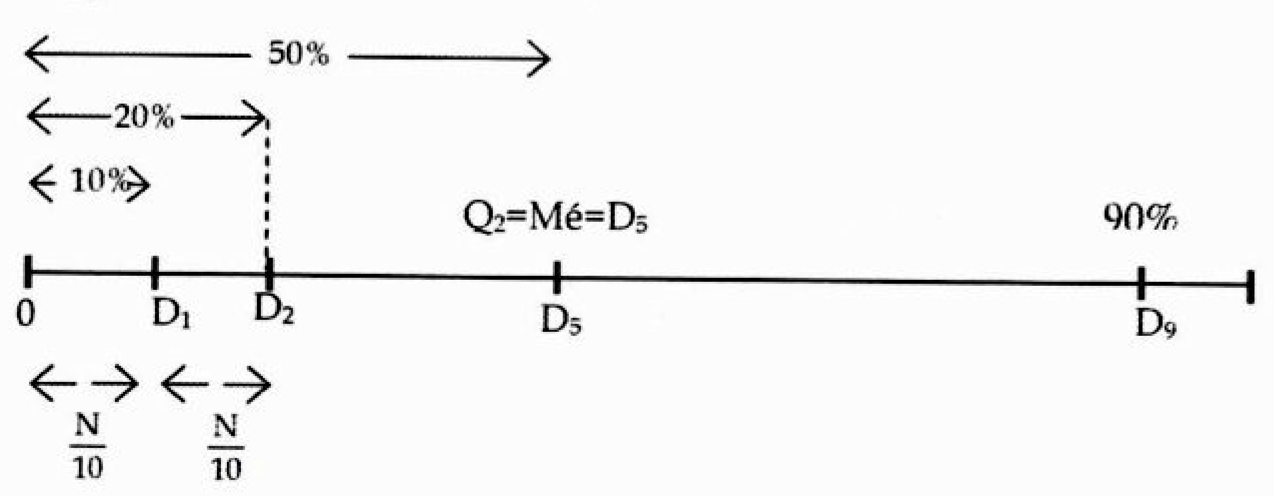
Pour comparer la dispersion entre deux séries statistiques ayant des unités différentes ou dont l'ordre de grandeur n'est pas le même, on utilise l'écart interquartile relatif donnée par :

$$\frac{Q_3 - Q_1}{Q_2}$$

3.- Les déciles :

Définition :

Les déciles sont les 9 valeurs $D_1, D_2, \ldots, D_8$ et D_9 de la variable statistique qui partagent la distribution en "10" parties composées du même effectif "$\frac{N}{10}$" et qui ont, donc, $N\frac{i}{10}$ observations inférieures à chaque D_i, i=1, 2,..., 9.



4.- Les percentiles :

***- Définition :**

Les percentiles sont les 99 valeurs $P_1, P_2, \ldots, P_{98}$ et P_{99} de la variable statistique qui partagent la distribution en "100" parties composées du même effectif "$\frac{N}{100}$" ayant, donc, N $\frac{i}{100}$ observations inférieures à chaque P_i, i=1, 2, ..., 99.

50%

2%

1%

$P_{50}=Q_2=Mé=D_5$

P_1 P_2 ... ... P_{98} P_{99}

$\frac{N}{100}$ $\frac{N}{100}$

***- Remarque :**

Le calcul des quantiles est pareil à celui de la médiane, en changeant la fréquence $\frac{N}{2}$ (nombre d'observations inférieures à la médiane) par $N\frac{i}{n}$ (nombre d'observations inférieures au quantile d'ordre $\frac{i}{n}$ cherché).

***- Exemple1 (cas du variable discrète "pondérée") :**

Calculer les percentiles 55 et 75 de la série statistique suivante :

x_i	n_i	$n_ic\uparrow$
3	3	3
4	7	10
8	30	40
10	20	60←
11	15	→75
20	25	100

N=100

1) $N\frac{55}{100} = 100 \cdot \frac{55}{100} = 55$

On cherche ce 55 entre les $n_ic\uparrow$

⇒ il n'existe pas exactement mais 60 est la 1ère valeur qui dépasse ce 55

alors $P_{55} = 10$

2) $N\frac{75}{100} = 75$, dans ce cas oui ⇒ $P_{75} = \frac{11+20}{2} = 15{,}5$

***- Exemple2 (cas du variable continue) :**

Calculons les déciles troisième et septième de la distribution statistique suivante :

$[e_{i-1}, e_i[$	n_i	$n_ic\uparrow$
[0, 10[	4	4
[10, 30[	8	12 *
[30, 35[	13	25
[35, 80[	5	30**
[80,100[	3	33
[100, 150[	7	40

N=40

$D_3 = ?$

$N\frac{3}{10} = 12$

cette valeur apparaît dans le tableau*, alors on prend $D_3 = e_2$ c'est-à-dire $D_3 = 30$

$D_7 = ?$

$$N\frac{7}{10} = 28$$

Cette valeur ne se trouve pas exactement dans le tableau, mais 30** est la 1ère valeur qui la dépasse, on applique alors la formule :

$$D_7 = e_{i-1} + \frac{N\frac{7}{10} - n_{i-1}c}{n_i} a_i$$

Qui vient de la formule générale :

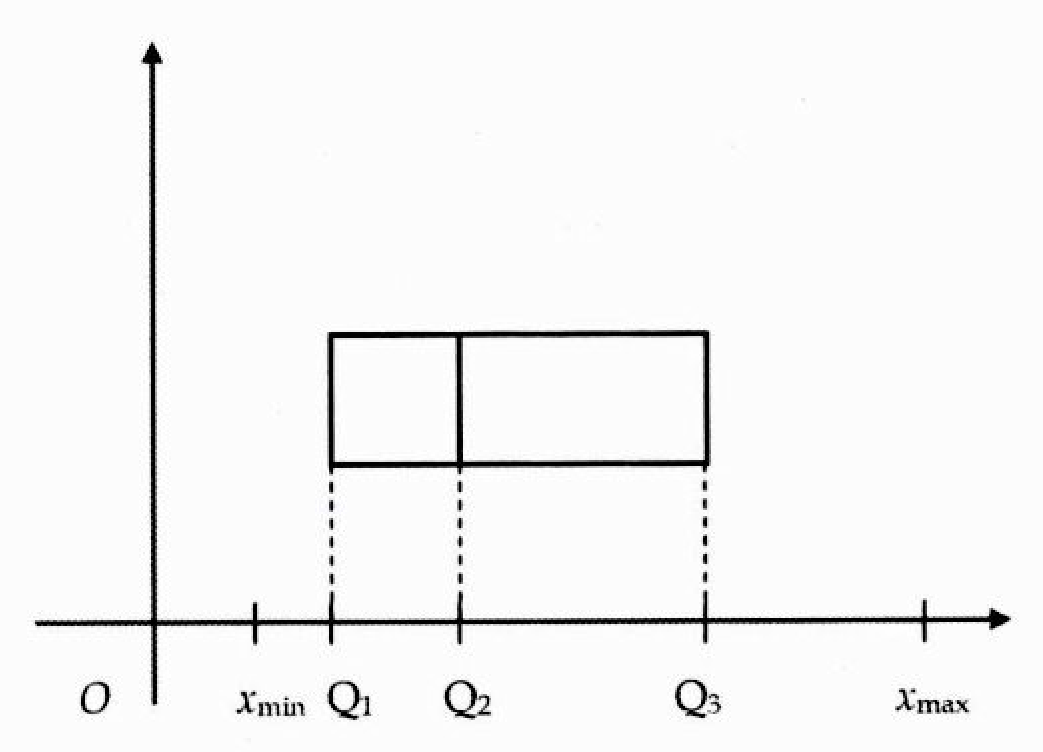
$$\text{Quantile} = e_{i-1} + \frac{N\frac{i}{n} - n_{i-1}c}{n_i} a_i$$

$$D_7 = 35 + \frac{28-25}{5}(80-35) = 35 + \frac{3}{5} \cdot 45 = 35 + 27$$

$$\boxed{D_7 = 62}$$

5.- Boîte de Tuckey. Diagramme de Box & Wiskers :

Considérons le diagramme en boîte ci-dessous, qui est la version la plus simple de la boîte de Tuckey, appliquée à la variable X.



On distingue sur ce schéma la « boîte de Tuckey » qui est le rectangle construit sur le premier quartile Q_1 et le troisième Q_3.

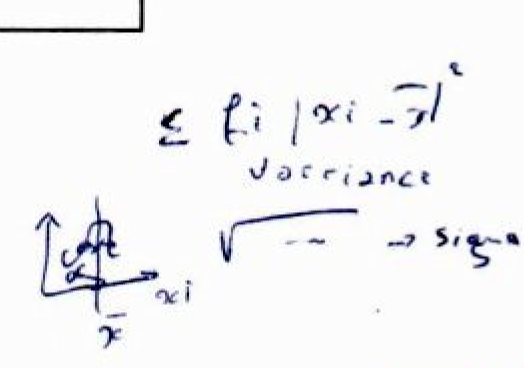
IV.- L'Ecart absolu moyen :

L'écart absolu moyen est la somme des valeurs absolues des écarts à la moyenne arithmétique, multipliées par les fréquences relatives.

$$e = \sum_{i=1}^{k} |x_i - \overline{x}| f_i$$

ou

$$\boxed{e = \frac{1}{N}\sum_{i=1}^{k} n_i |x_i - \overline{x}|}$$



Remarque :

Si l'écart absolu moyen d'une série statistique est grand (resp. petit), on dit que la série est très dispersée (resp. peu dispersée).

V.- La variance et l'écart-type :

- La variance d'une variable statistique X que l'on note Var(X) est la moyenne arithmétique des carrés des écarts des valeurs de la variable à leur moyenne arithmétique écarts des valeurs de la variable à leur moyenne arithmétique.

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \sum_{i=1}^{k}(x_i - \overline{x})^2 f_i$$

- L'écart-type d'une variable statistique X que l'on note σ_X est la racine carrée de la variance de X.

$$\sigma_X = \sqrt{Var(X)} = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2}$$

σ_X est donc la moyenne quadratique des écarts à la moyenne arithmétique.

***- Simplification des calculs de la variance :**

Pour simplifier le calcul de la variance, on peut utiliser un changement d'origine ou développer la formule de la façon suivante :

$$Var(X) = \sum_{i=1}^{k}(x_i - \overline{x})^2 f_i = \sum_{i=1}^{k}(x_i^2 - 2\overline{x}x_i + \overline{x}^2) f_i$$

$$Var(X) = \sum_{i=1}^{k} x_i^2 f_i - 2\overline{x}\underbrace{\sum_{i=1}^{k} x_i f_i}_{=\overline{x}} + \overline{x}^2\underbrace{\sum_{i=1}^{k} f_i}_{=1}$$

$$\Rightarrow \text{Var}(X) = \left(\sum_{i=1}^{k} x_i^2 f_i\right) - 2\overline{x}^2 + \overline{x}^2 = \left(\sum_{i=1}^{k} x_i^2 f_i\right) - \overline{x}^2$$

$$\text{Var}(X) = \left(\sum_{i=1}^{k} x_i^2 f_i\right) - \overline{x}^2 = \left(\frac{1}{N}\sum_{i=1}^{k} x_i^2 n_i\right) - \overline{x}^2$$

Le dit changement d'origine affecte la variance de la forme suivante :

$$x'_i = x_i - x_{i_o} \Rightarrow x'_i + x_{i_o} = x_i \Rightarrow \overline{x}' + x_{i_o} = \overline{x}$$

$$\text{Var}(X) = \sum_{i=1}^{k} (x_i - \overline{x})^2 f_i = \sum_{i=1}^{k} ((x_i^2 + x_{i_o}) - (\overline{x}' + x_{i_o}))^2 f_i$$

$$= \sum_{i=1}^{k} (x_i^2 - \overline{x}^2)^2 f_i = \text{Var}(X')$$

Donc le changement d'origine n'affecte pas la valeur de la variance.

Exemple :

Calculer la variance, en utilisant les simplifications des calculs, pour la série suivante :

$[e_{i-1}, e_i[$	n_i	c_i	c_i^2=ci-25	$n_i c_i^2$	$n_i c'^2_i$
[0, 10[	1	5	-20	-20	400
[10, 20[	2	15	-10	-20	200
[20, 30[	3	25	0	0	0
[30, 40[	4	35	10	40	400
	N=10			0	1000

On a pris $c_{i_0} = 25 = c_3$

$$Var(X^2) = \left(\frac{1}{N}\sum_{i=1}^{k} c_i'^2 n_i\right) - \overline{x}'^2 = \frac{1}{10}\cdot 1000 - \left(\frac{0}{10}\right)^2 = 100$$

$$\text{Donc} \quad Var(X) = 100$$

$$\text{et} \quad \sigma_X = \sqrt{Var(X)} = \sqrt{100} = 10$$

***- Variance intrasérie et variance intersérie :**

Supposons que l'on peut considérer deux sous-séries d'une série statistique. Alors il est possible de déterminer la variance globale de la série si l'on connaît les variances de ses deux sous-séries.

On a :

La moyenne de la série mère est la moyenne pondérée des moyennes des deux sous-séries :

$$\overline{X} = \frac{1}{N}(N_1\overline{x}_1 + N_2\overline{x}_2)$$

avec N est l'effectif total de la série mère

N_i et x_i sont l'effectif total et la moyenne arithmétique d'une sous-série.

La variance de la série total est égale à la moyenne des variances des deux sous-séries augmentée de la variance des moyennes des deux sous-séries.

$$\text{Var}(X) = \underbrace{\frac{1}{N}\left[N_1 \text{Var}(X_1) + N_2 \text{Var}(X_2)\right]}_{\substack{\text{Moyenne des variances} \\ \overline{\text{Var}(X_i)} \\ \text{Variance intrasérie}}} + \underbrace{\frac{1}{N}\left[N_1(\overline{x}_1 - \overline{x})^2 + N_2(\overline{x}_2 - \overline{x})^2\right]}_{\substack{\text{Variances des moyennes} \\ \text{Var}(\overline{X_i}) \\ \text{Variance intersérie}}}$$

Remarque :

- La variance intrasérie est la variance que l'on obtiendrait si les deux sous-séries avaient la même moyenne (égale à la moyenne globale $\overline{x}_1 = \overline{x}_2 = \overline{x}$).
- La variance intersérie est la variance que l'on obtiendrait si chaque variable de chaque sous-série était égal à sa moyenne (sous-séries homogènes).

Exemple :

Soit la série globale des salaires du personnel d'une société mère X composée de deux filiales X_1 et X_2.

	filiale X_1		filiale X_2		Société mère	
	effectif n_i	salaire x_i	effectif n_i	salaire x_i	effectif n_i	salaire x_i
Cadres	10	80	5	70	15	76,67
Techniciens	20	18	10	16	30	17,33
Ouvriers	30	10	100	8	130	8,46
	N1 = 60	$\overline{X}_1 = 24{,}33$	N2 = 115	$\overline{X}_2 = 11.39$	N = 175	$\overline{X} = 15{,}83$

Calculons la variance globale, variance intrafiliale et variance interfiliale.

Variance globale :

$$\mathrm{Var}(X) = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right) - \overline{x}^2$$

$$= \frac{1}{175}((15 \times 76{,}67^2) + (30 \times 17{,}33^2) + (130 \times 8{,}46^2)) - (15{,}83)^2$$

$\mathrm{Var}(X) \approx 359$

Variance intrafiliale :

$$\overline{\mathrm{Var}(X_i)} = \frac{1}{N}[N_1 \mathrm{Var}(X_1) + N_2 \mathrm{Var}(X_2)]$$

avec $\mathrm{Var}(X_1) = \left(\frac{1}{N_1}\sum_{i=1}^{k} n_i x_i^2\right) - \overline{x}^2$

$$\mathrm{Var}(X_1) = \frac{1}{60}\left(\left(10 \times 80^2\right) + \left(20 \times 18^2\right) + \left(30 \times 10^2\right)\right) - 24{,}33^2 = 632{,}72$$

$$\mathrm{Var}(X_2) = \left(\frac{1}{N_2}\sum_{i=1}^{k} n_i x_i^2\right) - \overline{x}^2$$

$$\mathrm{Var}(X_2) = \frac{1}{115}\left(\left(5 \times 70^2\right) + \left(10 \times 16^2\right) + \left(100 \times 8^2\right)\right) - 11{,}39^2 = 161{,}23$$

ainsi $\overline{\mathrm{Var}(X_i)} = \frac{1}{175}[(60 \times 632{,}72) + (115 \times 161.23)] \approx 322$

Variance interfiliale :

$$\mathrm{Var}(\overline{X}_i) = \frac{1}{N}\left[N_1(\overline{X}_1 - \overline{X})^2 + N_2(\overline{X}_2 - \overline{X})^2\right]$$

$$\mathrm{Var}(\overline{X}_i) = \frac{1}{175}\left[60\,(24{,}33 - 15{,}83)^2 + 115(11{,}39 - 15{,}83)^2\right]$$

$$\mathrm{Var}(\overline{X}_i) \approx 37$$

Conclusion :

$$\mathrm{Var}(X)=\overline{\mathrm{Var}(X_i)}+\mathrm{Var}(\overline{X_i})$$

$$359 = 322 + 37$$

VI.- Le coefficient de variation :

Il est égal au rapport de l'écart-type à la moyenne et il est présenté sous forme de nombre abstrait sans dimension et indépendant des unités de mesure.

$$C_V=\frac{\sigma(X)}{\overline{X}}$$

Le coefficient de variation permet de donner une idée sur l'amplitude de variation en comparant l'écart-type et la moyenne. Ce rapport est généralement exprimé en pourcentage.

$$C_V=\frac{\sigma(X)}{\overline{X}}\times 100$$

Plus le coefficient de variation est petit, plus la série est homogène. D'une manière générale, la population étudiée est considérée homogène lorsque le Cv < 30%.

Exemple :

Calculer C_V pour l'exemple des deux filiales X_1 et X_2 précédents :

$$C_V(X_1)=\frac{\sigma(X_1)}{\overline{X}_1}=\frac{12,7}{11,39}=1,12$$

$$C_V(X_2)=\frac{\sigma(X_2)}{\overline{X}_2}=\frac{25,15}{24,33}=1,03$$

la dispersion relative des salaires et presque identique dans les deux établissements.

Exercices

Exercice n°1 :

Les salaires annuels (en 10^3 DH) de 100 ouvriers d'une entreprise sont distribués de la manière suivante :

Salaires (en 10^3 DH) compris entre	Nombres des ouvriers
25 et 26	13
26 et 28	15
28 et 30	21
30 et 32	19
32 et 34	15
34 et 35	10
35 et 36	7

$\frac{30+32}{2} = 31$

$N = 100$

1) Calculer les fréquences relatives et les effectifs cumulés croissants.
2) Déterminer le nombre des ouvriers qui ont un salaire annuel inférieur à 32000 DH.
3) Déterminer le pourcentage des ouvriers qui gagnent moins que 28350 DH par an.
4) Déterminer le salaire annuel le plus fréquent graphiquement et par le calcul.
5) Déterminer la médiane de cette distribution et interpréter le résultat.
6) Calculer le salaire annuel moyen de ces ouvriers.

7) Calculer les quartiles Q_1 et Q_3 de cette distribution.

8) Déterminer sa variance et son écart-type.

Exercice n°2:

Une étude sur l'épargne locale auprès des agences relevant des délégations provinciales de la Poste a permis d'établir les caractéristiques des versements à la caisse d'épargne, au cours d'une journée type, dans la région Tanger-Tétouan. Ces caractéristiques sont résumées dans le tableau ci-après :

Etat de versement à la Caisse d'épargne	Tanger	Tétouan
Nombre de versements	35	30
Versements moyens	2200	2400
Variances des versements	3025	3600

1) Donner le versement moyen pour l'ensemble de la région Tanger-Tétouan.
2) Calculer la variance intra-ville et la variance inter-ville des versements.

   Quelle est la variance régionale des versements ?
3) Comparer la dispersion des versements à Tanger et Tétouan, en utilisant le coefficient de variation.

Exercice n°3 : ✗

Une enquête sur la mobilité a donné la répartition suivante (exprimée en %) pour une population d'individus domiciliés à la région Tanger-Tétouan, selon la distance entre le domicile et le lieu de travail (distance exprimée en kilomètres) :

Distance (en km)	Tétouanais (en %)	Tangérois (en %)
[0, 2[	5.7	3.1
[2, 6[	12.0	7.7
[6, 10[	6.3	6.8
[10, 15[	8.1	4.6
[15, 25[	11.0	6.8
[25, 50[	11.2	5.4
[50, 60[	7.8	3.5

1. Calculez la distance moyenne parcourue par un travailleur tétouanais et par un tangérois.
2. Déterminez la variance et l'écart-type des distances parcourues par un travailleur tétouanais et par un tangérois.
3. Calculer le coefficient de variation pour les distances parcourues par un travailleur tétouanais et pour celles parcourue par un tangérois. Conclure.

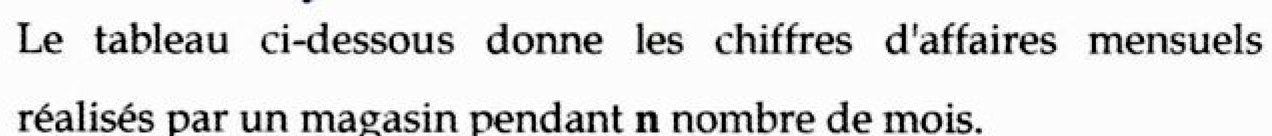
Exercice n°4 :

Les salaires annuels (en 1000 DH) des employés d'une entreprise composée de deux filiales X et Y sont réparties selon le tableau suivant :

Salaires en 1000 DH compris entre	Nombre d'employés de la filiale X	Nombre d'employés de la filiale Y
10 et 20	5	4
20 et 30	10	12
30 et 40	13	14
40 et 50	4	6

1) Déterminer le salaire moyen de la filiale X et de la filiale Y.
2) Déterminer la variance et l'écart-type des salaires de la filiale X et de la filiale Y.
3) Comparer la dispersion des salaires de la filiale X et de la filiale Y.

Exercice n°5 :



Le tableau ci-dessous donne les chiffres d'affaires mensuels réalisés par un magasin pendant **n** nombre de mois.

Chiffre d'affaires (en 10^6 DH)	Nombre de mois (ni)
[3, 4[	3
[4, 6[	4
[6, 8[	n_3 8
[8, 10[	5
[10, 14[	n_5 5.
[14, 18[	6

[18, 24[	6
[24, 30[	13

On donne en plus les informations suivantes:

-la densité d'effectif de la 3ème classe est le double de la deuxième

-la densité d'effectif de la cinquième classe est identique à celle de la quatrième.

1) Définir pour cette distribution : la population statistique, l'unité statistique, le caractère et la nature de la variable statistique.

2) Déterminer $\mathbf{n_3}$, $\mathbf{n_5}$ et $\mathbf{n}$.

3) Tracer l'histogramme de cette distribution. On présentera au préalable un tableau précisant pour chaque classe, son amplitude $\mathbf{a_i}$ et la hauteur de chaque classe $\mathbf{h_i}$ (sans utiliser les fréquences) représentée sur l'histogramme. En déduire sa classe modale en précisant sa signification.

4) Calculer le chiffre d'affaires mensuelles médianes Me par interpolation linéaire, en précisant l'hypothèse sous-jacente à cette méthode.

5) On regroupe la 4ème et la 5ème classe ensemble. Calculer dans ce cas le chiffre d'affaires mensuels moyen $\bar{x}$ et l'écart type σ de la variable étudiée.

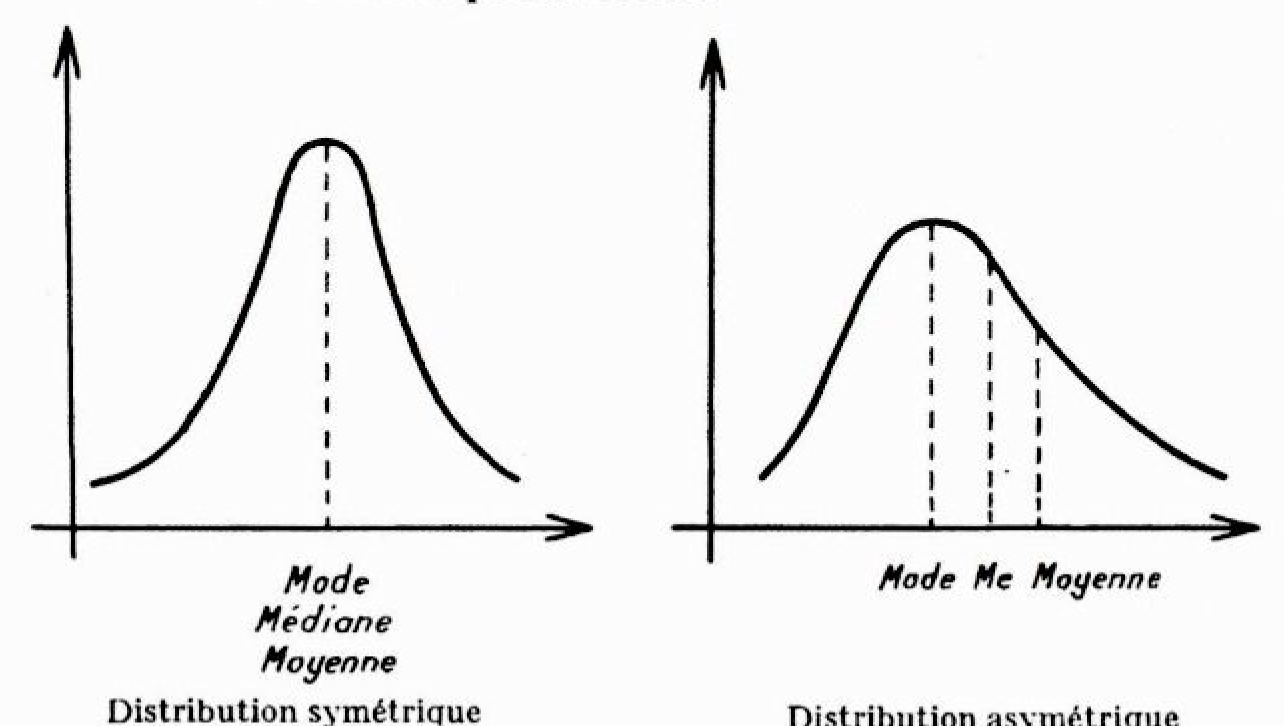
6) En utilisant ces résultats calculer le pourcentage des observations contenues dans un intervalle symétrique I de longueur 2σ et centré en $\bar{x}$ ($I = [\bar{x} - \sigma, \bar{x} + \sigma]$)

Chapitre 4 :

Caractéristiques de forme

I.- Introduction :

Lorsque l'on représente graphiquement une série statistique, on peut remarquer qu'une telle distribution présente une symétrie ou non, mais pour la mesurer, par exemple, on doit utiliser une caractéristique de forme.



Distribution symétrique

Distribution asymétrique

II.- Les Moments:

1.-Moment simple :

Le moment d'ordre r d'une série statistique par rapport à une valeur quelconque x_o est :

$$m_r=\frac{1}{N}\sum_{i=1}^{k}(x_i-x_o)^r n_i$$

si $x_o = 0$ alors $m_r=\frac{1}{N}\sum_{i=1}^{k}x_i^r n_i$

ce qui donne le moment simple d'ordre r.

Remarques :

* La moyenne arithmétique

$\overline{X}=\frac{1}{N}\sum_{i=1}^{k}x_i n_i = m_1$ est le moment simple d'ordre 1.

* Le moment simple d'ordre 0 est

$$m_0=\frac{1}{N}\sum_{i=1}^{k}x_i^0 n_i=\frac{1}{N}\sum_{i=1}^{k}n_i=\frac{N}{N}=1$$

2.-Moment centré :

Lorsque $x_o=\overline{x}$, on obtient le moment centré d'ordre r :

$$\mu_r=\frac{1}{N}\sum_{i=1}^{k}(x_i-\overline{x})^r n_i$$

Si r est pair, μ_r est un paramètre de dispersion.

Si r est impair, μ_r est un paramètre de symétrie.

En particulier :

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \overline{x})n_i$$

$$= \frac{1}{N}\sum_{i=1}^{k}x_i n_i - \frac{1}{N}\overline{x}\sum_{i=1}^{k}n_i = \overline{x} - \overline{x} = 0$$

variation

et $\mu_2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = \mathrm{Var}(X)$.

3.- Relation entre mr et μr :

En utilisant que ∀x et ∀y des réels

$$(x + y)^r = \sum_{i=1}^{r} C_r^i x^{r-i} y^i$$ Formule de binôme de Newton

ou les identités remarquable $(a - b)^2$, $(a + b)^2$,..., on obtient :

var(X) = (1/N Σ ni xi) − x̄² → Formule Réduite de la variance

la dispersion de

$$\mu_2 = m_2 - m_1^2$$
$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3$$
$$\mu_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4$$
$$\mu_5 = m_5 - 5m_1m_4 + 10m_1^2m_3 - 4m_1^5$$

...etc.

III.- L'Asymétrie

Une distribution statistique est symétrique si sa représentation graphique est symétrique par rapport à l'axe vertical $x = \overline{x}$ ou si les valeurs équidistantes de sa valeur centrale ont une fréquence égale.

Dans ce point central, lorsque la série est symétrique, coïncident la moyenne arithmétique, la médiane et le mode.

$$\overline{x} = Mé = Mo$$

(s'il y a plus d'une seule mode, coïncident seulement $\overline{x} = Mé$).

Exemple1 :

x_i	n_i	$n_ic\uparrow$	n_ix_i
3	3	3	9
5	2	5	10
7	4	9*	28
9	2	11	18
11	3	14	33
	N=14		98

$$\overline{x} = \frac{98}{14} = 7$$

$$\frac{N}{2} = \frac{14}{2} = 7 \Rightarrow Mé = 7$$

$$Mo = 7$$

$$\overline{x} = Mé = Mo = 7$$

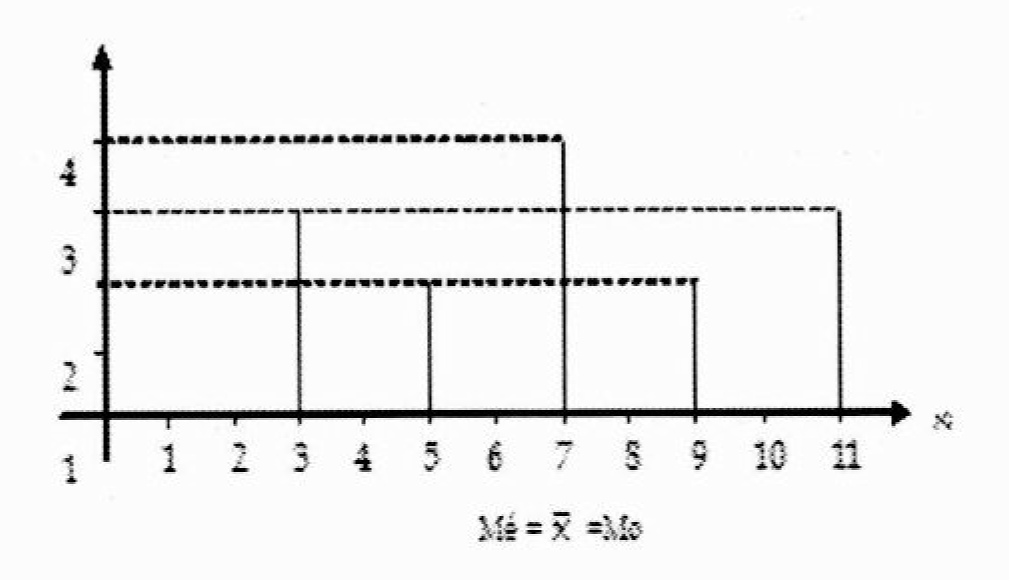


Diagramme en batôns des n_i

Exemple2 :

x_i	n_i	$n_ic\uparrow$	n_ix_i
3	5	5	15
5	1	6	5
7	3	9*	21
9	1	10	9
11	5	15	55
	N=15		105

$$\overline{x} = \frac{105}{15} = 7$$

$$\frac{N}{2} = \frac{15}{2} = 7{,}5 \Rightarrow Mé = 7$$

Mo = 3 et 11

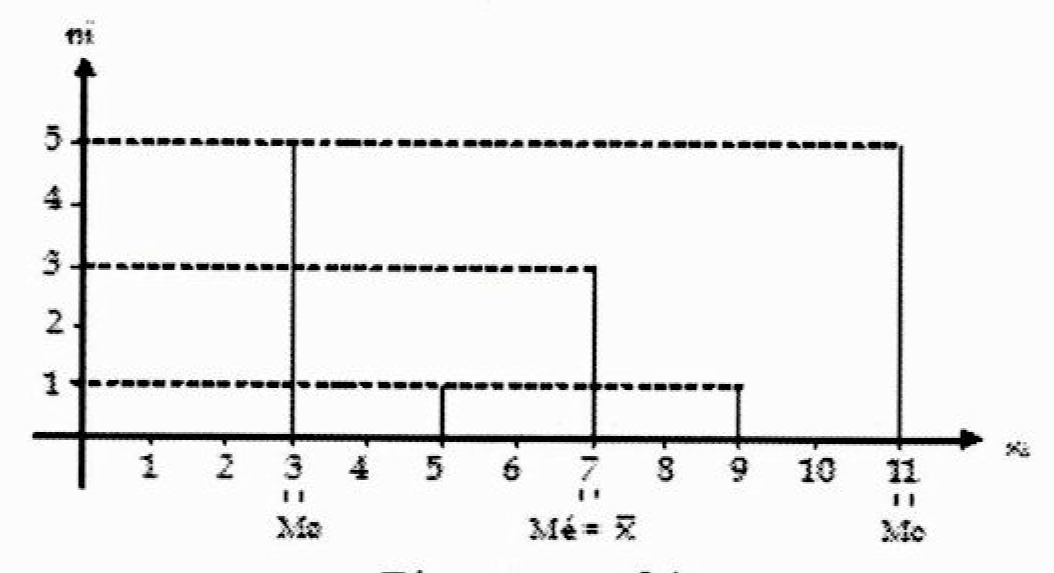


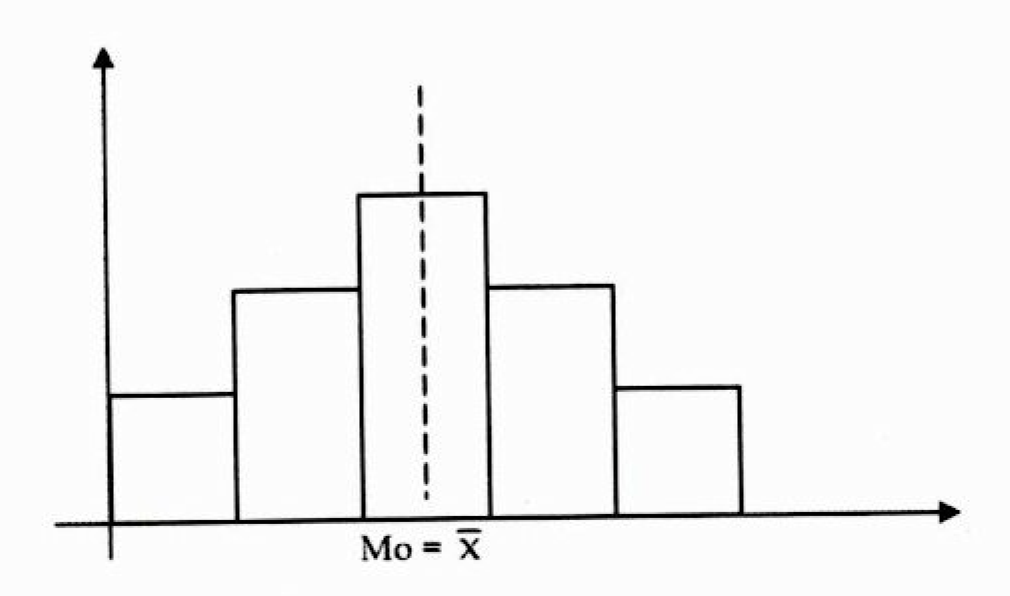
Diagramme en bâtons

Lorsque la série statistique n'est pas n'est pas symétrique, on dit qu'elle est asymétrique et alors on a plus $\overline{x}$ = Mé =Mo.

On peut mesurer son asymétrie par le coefficient de Pearson qui se définit par :

$$A_P = \frac{\overline{x} - Mo}{\sigma(x)}$$

En effet, si la série est symétrique et elle a un seul mode, alors $\overline{x}$ =Mo et A_P=0.



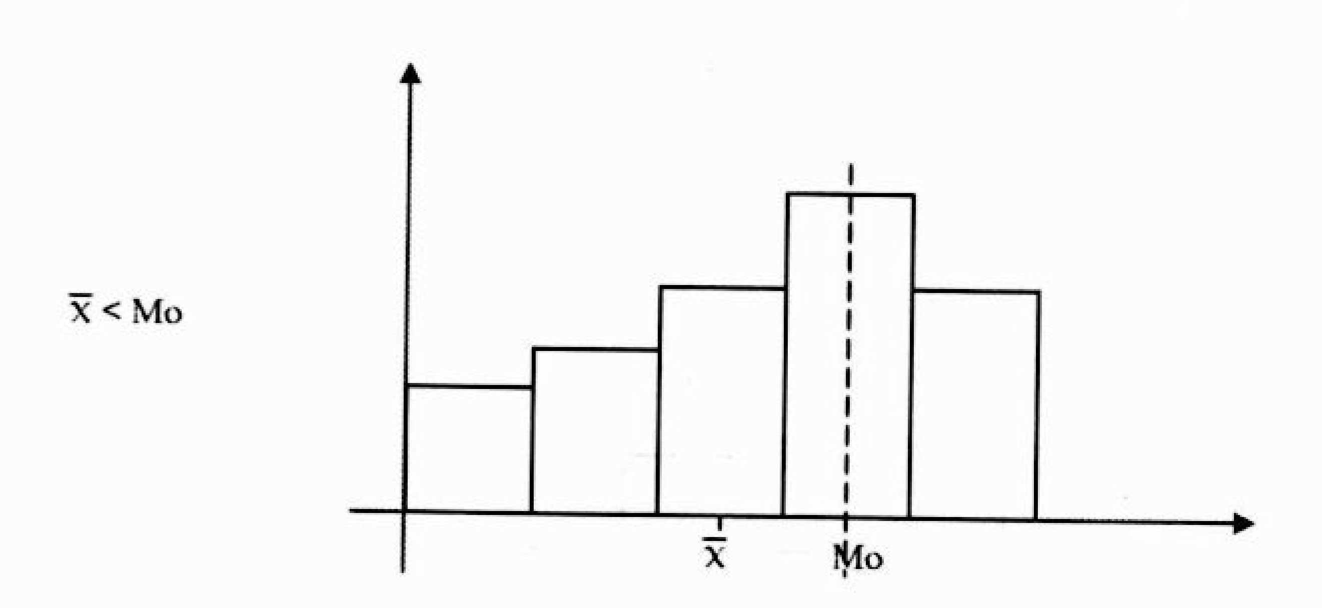
Le coefficient d'asymétrie le plus important et le plus utilisé est le coefficient de Fisher que se définie comme :

$$g_1 = \frac{\mu_3}{\sigma(x)^3}$$

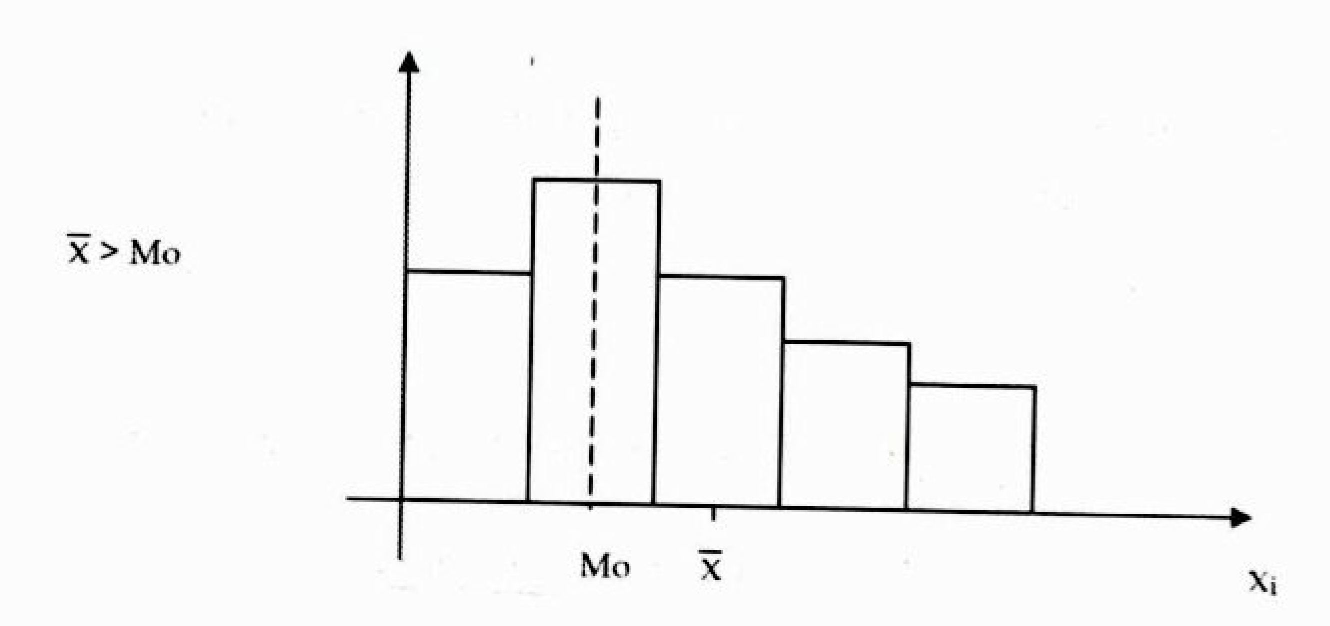
la forme μ_3

avec $\mu_3 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \overline{x})^3 n_i$ est le moment centré d'ordre 3.

- si $g_1 = 0$ alors la distribution est symétrique.
- si $g_1 < 0$ alors la distribution est asymétrique à gauche ou on dit qu'elle est oblique à droite.



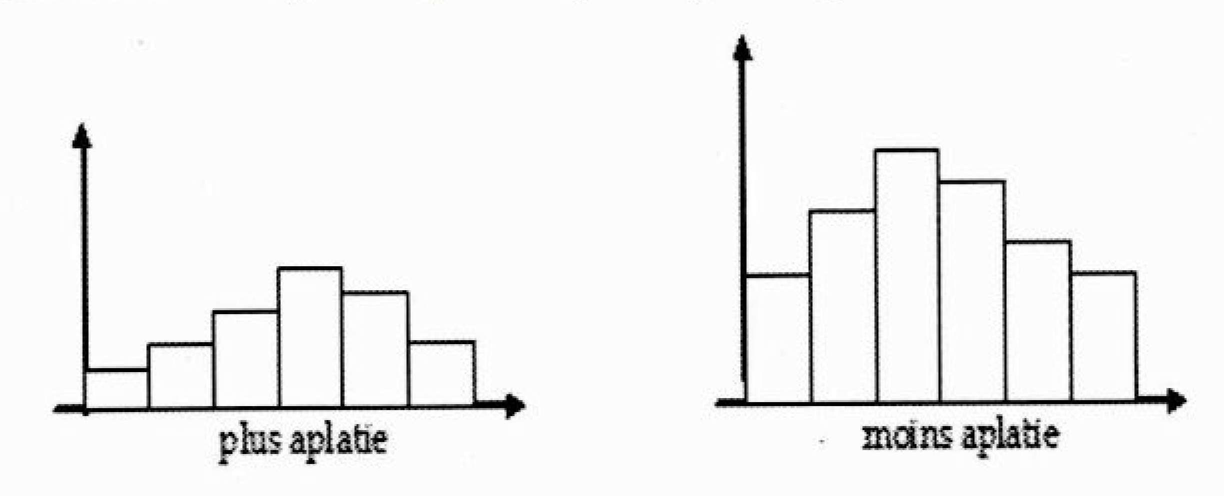
- si $g_1 > 0$ alors la distribution est asymétrique à droite ou on dit qu'elle est oblique à gauche.



IV.- L'Aplatissement

Une série statistique est plus ou moins aplatie suivant que les fréquences des valeurs proches des valeurs centrales sont plus au moins élevées par rapport aux autres.

Les mesures d'aplatissement s'appliquent aux distributions unimodales et symétriques ou peu asymétriques.



Entre les deux cas, il existe un cas normal ou moyen c'est le cas des distributions moyennement aplaties.

- Les séries statistiques plus aplaties sont dites PLATYKURTIQUES.
- Les séries statistiques moins aplaties sont dites LEPTOKURTIQUES.
- Les séries statistiques moyennement aplaties sont dites MÉSOKURTIQUES.

"PLATUS" signifie large.

"LEPTOS" signifie mince.

"MÉSO" signifie moins que, inférieur à, manque de...

"KURTOSIS" signifie sommet ou bosse.

Pour mesurer le degré d'aplatissement, on utilise le coefficient de Kurtosis.

$$g_2 = \frac{\mu_4}{\sigma(x)^4} - 3$$

- si $g_2 = 0$ alors la distribution est mésokurtique ou **normale**.
- si $g_2 < 0$ alors la distribution est platykurtique.
- si $g_2 > 0$ alors la distribution est leptokurtique.

Exercices

Exercice n°1 :

Le tableau ci-dessous donne les chiffres d'affaires mensuels réalisés par une entreprise commerciale pendant 36 mois :

Chiffres d'affaires (en 106 DH)	Nombre de mois (ni)
[3, 6[	7
[6, 8[	6
[8, 10[	5
[10, 14[	8
[14, 18[	6
[18, 23[	4

1°) Calculer le mode et la moyenne de cette série statistique.

Cette distribution statistique est-elle symétrique ? Pourquoi ?

2°) Calculer sa variance et son écart type.

3°) Déterminer son coefficient de variation. **Conclure.**

Exercice n°2 :

On étudie les revenus (Annuels en milliers de dirhams) d'un ensemble de familles d'un quartier de Tétouan, les données sont regroupées dans le tableau suivant :

Revenus annuels (en 10^3 DH)	[18 ; 30[	[30; 36[	[36 ; 42[	[42 ; 54[	[54 ; 60[	[60 ; 66[
Effectifs	13	219	20	46	50	82

1. Préciser les caractéristiques de cette série (population, taille ou l'effectif total, individu, caractère étudié, type de caractère et modalités).
2. Calculer la moyenne $\overline{x}$ de cette série statistique
3. Dresser l'histogramme de cette série statistique puis représenter son polygone.
4. Déterminer le mode *Mo* de cette série, graphiquement et par le calcul.
5. Calculer la médiane *Mé* de cette série statistique en explicitant vos calculs.
6. La série étudiée est-elle symétrique ou asymétrique ? Justifier votre réponse. Pouvait-on prévoir ce résultat?

Chapitre 5 :

Caractéristiques de concentration

I.- Introduction :

Plusieurs phénomènes économiques nécessitent des études de concentration et répartition d'une des variables dont ils dépendent. Par exemple, la concentration des salaires, des revenus, de production,...etc. Les caractéristiques déjà vues comme la moyenne, la variance, etc... ne répondent pas à cette étude, alors on a construit de nouvelles caractéristiques dites de concentration mais due à la nature des problèmes étudiés, on considère que des variables continues à valeurs positives.

II.- La médiale (Ml) :

La médiale est la valeur de la variable statistique telle que la somme des observations inférieure à elle est égale à la somme des observations supérieure à elle. Donc, c'est une médiane sur les valeurs "n_ic_i" au lieu des "n_i". Pour cela la médiale se calcul de forme analogue à la médiane.

Exemple :

Calculons la médiale des salaires annuels du personnel (en milliers de DH) d'une entrepris :

$[e_{i-1}, e_i[$	n_i	c_i	n_ic_i	$(n_ic_i)c\uparrow$
[40, 50[	12	45	540	540
[50, 60[	14	55	770	1310
[60, 70[	20	65	1300	2610
[70, 80[	30	75	2250	4860 *
[80, 90[	14	85	1190	6050
[90, 100[	10	95	950	7000
	N=100		7000	

On a $\frac{\sum n_i c_i}{2} = \frac{7000}{2} = 3500$. On cherche cette valeur dans la colonne des $(n_ic_i)c\uparrow$. la 1ère valeur qui la dépasse est 4860. Donc la classe médiale est [70, 80[

On applique, alors, une formule analogue à celle de calcul de la médiane :

$$Ml = l_{i-1} + \frac{\frac{\sum n_i c_i}{2} - (n_{i-1}c_{i-1})c\uparrow}{n_i c_i} \cdot a_i$$

$$Ml = 70 + \frac{3500 - 2610}{2250} \cdot 10 \approx 73{,}96$$

Interprétation :

La masse salariale de l'entreprise, versée annuellement aux employés montre que le partage en 2 blocs égaux s'effectue pour la valeur de 73.960 DH. Autrement dit, la moitié de la masse salariale est reçue par le personnel qui a un salaire inférieur à

73.960 DH et l'autre moitié est reçue par ceux qui ont un salaire supérieur à 73.960 DH.

Calcul de la concentration :

Il existe plusieurs démarches pour effectuer ce calcul, la plus simple se compose des étapes suivantes :

- Calcul de la médiane de la série.
- Calcul de la médiale de la série.
- Mesure de l'écart médiale-médiane (ΔM).
- Comparaison de cet écart à l'intervalle de variation de la série.

Calcul de la concentration pour l'exemple précédent :

On a déjà calculé, pour cet exemple, la médiane et la médiale, on a :

$$Mé = 71{,}33 \qquad \text{et} \qquad Ml = 73{,}96$$

$$\Delta M = Ml - Mé = 73{,}96 - 71{,}33 = 2{,}63.$$

L'intervalle de variation est : 100 – 40 = 60, d'où

$$I = \frac{\Delta M}{\text{intervalledevariation}} = \frac{2{,}63}{60} = 0{,}043$$

En générale, l'indice $I = \dfrac{\Delta M}{\text{intervalledevariation}} = \dfrac{Ml - Mé}{\text{étendue}}$ est une valeur comprise entre 0 et 1 qui mesure la concentration, indiquant une grande concentration lorsqu'elle est grande (proche

de 1) et une équi-distribution lorsqu'elle est petite (proche de zéro).

Si I = 0, alors il y a équi-distribution absolue.

Si I = 1, alors il y a maximum concentration.

Donc, pour l'exemple, les salaires sont équi-distribués. On a la concentration est faible.

III.- Courbe de Lorenz et Indice de Gini:

1.- Courbe de Lorenz :

Considérons, le cas de l'étude de la variable qui représente les salaires d'un groupe d'ouvriers. On a construit le tableau suivant pour analyser la concentration des salaires :

$[e_{i-1}, e_i[$	n_i	c_i	n_ic_i	$n_ic\uparrow$	$(n_ic_i)c\uparrow$	$p_i = \frac{n_ic\uparrow}{N}\cdot 100$	$q_i = \frac{(n_ic_i)c\uparrow}{\sum_i n_ic_i}\cdot 100$
$[e_0, e_1[$	n_1	c_1	n_1c_1	$n_1c\uparrow$	$(n_1c_1)c\uparrow$	p_1	q_1
$[e_1, e_2[$	n_2	c_2	n_2c_2	$n_2c\uparrow$	$(n_2c_2)c\uparrow$	p_2	q_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
$[e_{i-1}, e_i[$	n_i	c_i	n_ic_i	$n_ic\uparrow$	$(n_ic_i)c\uparrow$	p_i	q_i
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
$[e_{k-1}, e_k[$	n_k	c_k	n_kc_k	$n_kc\uparrow$	$(n_kc_k)c\uparrow$	p_k	q_k
	N		$\sum_i n_ic_i$				

Où les produits "n_ic_i" représentent la masse salariale reçue par les "n_i" ouvriers dont le salaire est compris entre e_{i-1} et e_i.

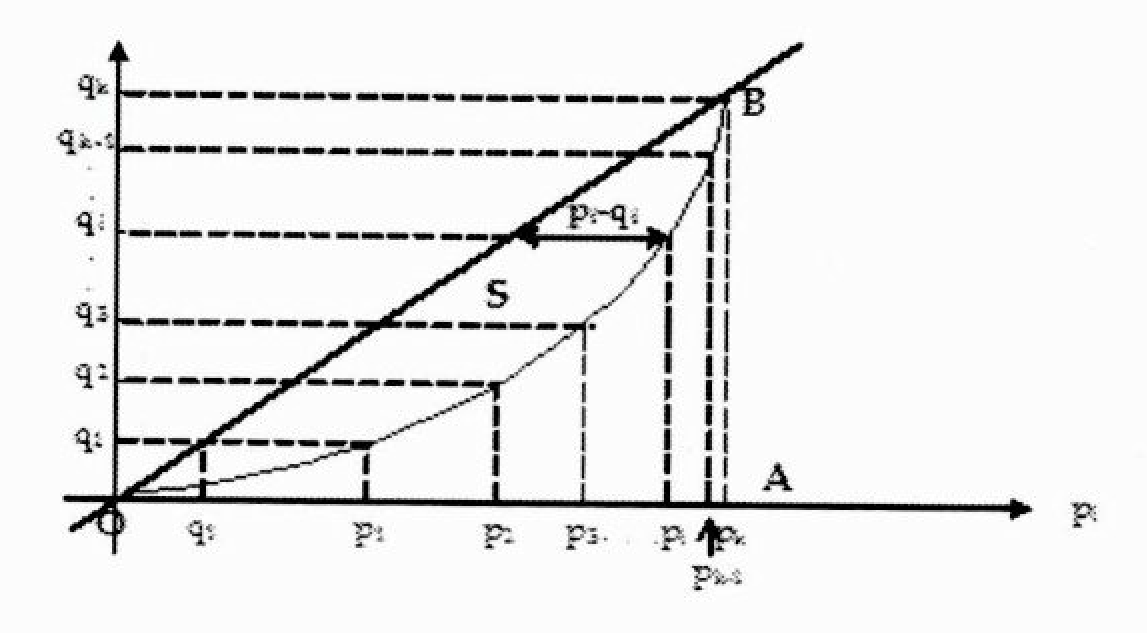
Les "$(n_ic_i)c\uparrow$" représentent la masse salariale reçue par les "$n_ic\uparrow$" ouvriers avec des salaires inférieurs à cette masse.

Les valeurs "p_i" représentent, en pourcentage, les fréquences relatives cumulées des ouvriers.

Les quantités "q_i" représentent la masse salariale cumulée (exprimées en pourcentage) sur la masse salariale totale $\sum_i n_i c_i$.

Les deux dernières colonnes de tableau nous informent sur la concentration des salaires.

La distribution des ouvriers et des salaires exprimée par les valeurs (p_i, q_i); i=1, ..., k, se représente graphiquement à l'aide de la courbe de concentration appelée courbe de Lorenz ; c'est la courbe qui passe par les points (p_i, q_i)



En gras c'est la première bissectrice.

- Si les salaires sont équi-distribués, alors $p_i=q_i$ ($\forall i=1, \ldots, k$), donc la courbe de Lorenz coïncide avec la 1ère bissectrice.
- Dans le cas de maximum concentration $q_i=0$ ($i=1, \ldots, k-1$) et $qk=100$, alors la courbe de concentration coïncide pratiquement avec le triangle OAB.
- Pour des situations de concentration intermédiaires, la courbe de Lorenz occupe une position intermédiaire entre ces situations extrêmes. Elle s'approche plus à la 1ère bissectrice lorsque la concentration est faible et s'éloigne de la 1ère bissectrice lorsque la concentration est forte.

2.- Indice de Gini

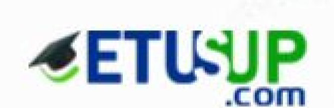
A partir de la courbe de Lorenz, on construit une mesure de degré de concentration appelée "Indice de Gini" notée "I_G".

Géométriquement,

$$I_G = \frac{\text{l'aire S}}{\text{l'aire du triangleOAB}}$$

$$\text{et on a : } 0 \leq I_G \leq 1$$

Evidement, cet indice est une mesure de l'approximation de la courbe de Lorenz à la 1ère bissectrice et par suite de la concentration des salaires.



Algébriquement,

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i}$$

qui mesure aussi l'approximation de la courbe de Lorenz à la 1[ère] bissectrice mais en utilisant une autre propriété géométrique différente de celle de l'aire S.

- Si $I_G = 0$ alors pi=qi et les salaires sont équi-distribués.
- Si $I_G = 1$ alors qi = 0 et il y a maximum concentration.
- Si $0 < I_G < 1$ alors il y a grande concentration si I_G est plus proche de 1 ; et une équi-distribution si I_G est plus proche de 0.

Exemple :

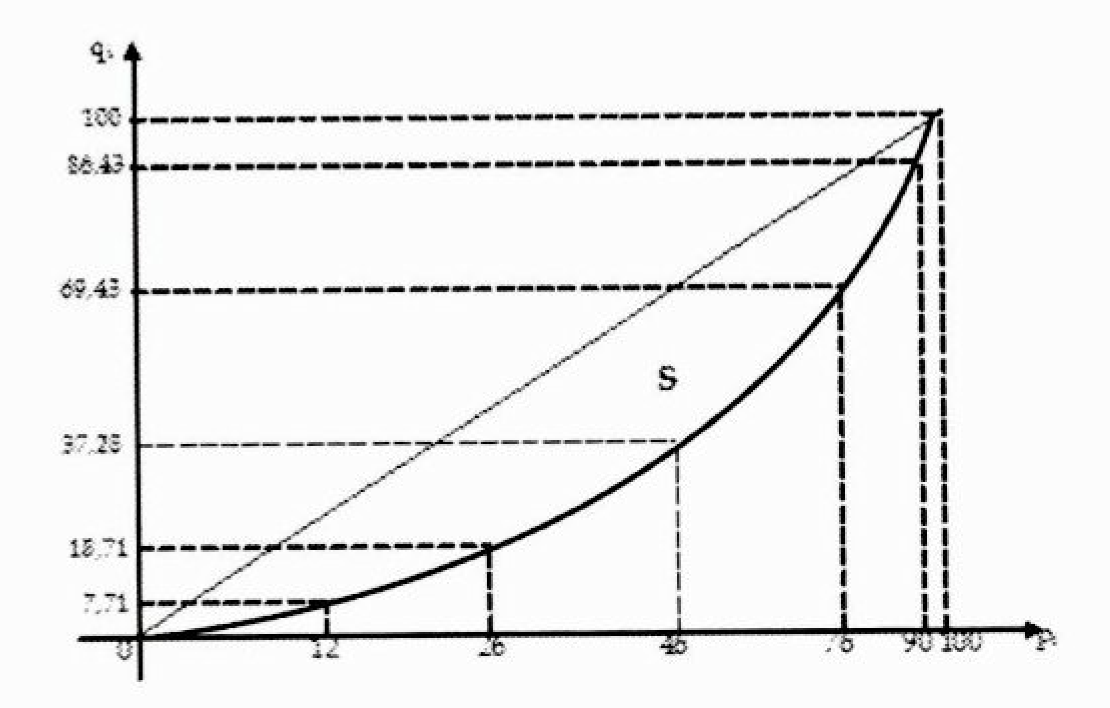
Reprenons l'exemple précèdent ; on a calculé l'indice I pour cet exemple et on a trouvé I = 0,043.

Calculons, maintenant l'indice de Gini I_G.

$[e_{i-1}, e_i[$	n_i	c_i	n_ic_i	$n_ic\uparrow$	$(n_ic_i)c\uparrow$	$p_i = \frac{n_ic\uparrow}{N} \cdot 100$	$q_i = \frac{(n_ic_i)c\uparrow}{\sum_i n_ic_i} \cdot 100$
[40, 50[	12	45	540	12	(540	12	7,71
[50, 60[	14	55	770	26	1310	26	18,71
[60, 70[	20	65	1300	46	2610	46	37,28
[70, 80[	30	75	2250	76	4860	76	69,43
[80, 90[	14	85	1190	90	6050	90	86,43
[90, 100[	10	95	950	100	7000	100	100
	N=100		7000				

Si nous observons les deux dernières colonnes nous avons :

Au 12% des ouvriers qui gagnent peu leur correspondent le 7,71% du total des salaires, au 26% des ouvriers qui gagnent peu leur correspond le 18,71% du total des salaires, etc... Alors que s'il avait une répartition égalitaire des salaires, nous aurons que au 12% des ouvriers leur correspond le 12% du total des salaires, au 26% des ouvriers leur correspond le 26% du total des salaires, etc.... ce que l'on n'a pas dans cet exemple du à une telle concentration des salaires (même fiable !).



L'indice de Gini

$$I_G = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{219,56}{250}$$

$$I_G = 1 - 0,878 = 0,122$$

Ce qui indique une faible concentration des salaires de ce groupe des ouvriers.

Exercices

Exercice n°1:

Soit la série statistique des salaires d'une entreprise :

Salaires	Nombre d'employés
X – 50	30
50 – 100	40
100 – 200	20
200 – 300	10

1) Retrouver la borne inférieure X de la première classe sachant que le salaire moyen est de 94.

(Pour la suite des calculs, retenez la valeur trouvée à la première question.)

2) Donner l'interprétation et la valeur de la médiane (Mé).
3) Calculer le troisième quartile, le septième décile et le percentile 35.
4) Déterminer la variance et l'écart-type.
5) Donner l'interprétation et la valeur de la médiale (Ml).
6) Que peut-on dire de la différence ΔM=Ml-Mé ? Comparer-la à l'étendue. Interpréter le résultat.
7) Construire la courbe de Lorenz.

8) Déterminer l'indice de concentration de Gini. Conclure.

Exercice n°2 :

Pour une même durée de travail, les salaires d'une entreprise se répartissent comme suit :

Salaires en dirhams	Nombre de personnes
De 3000 à 4000	11
De 4000 à 5000	26
De 5000 à 6000	63
De 6000 à 7000	81
De 7000 à 8000	35
De 8000 à 9000	21
De 9000 à 10 000	13

1°) Déterminer le salaire médian.

2°) Déterminer la médiale.

3°) Tracer la courbe de concentration de Lorenz et calculer l'indice de Gini. **Conclure.**

Exercice n° 3:

La distribution des salaires horaires, en dirhams, des **N** employés d'une grande entreprise est donnée par :

Classes	Effectifs
[50 ; 100[	10
[100 ; 150[	14
[150 ; 200[	16
[200 ; 250[	**n**

Ces données sont incomplet car, à la suite d'un incident, l'effectif de la dernière classe est illisible ; alors, on a décidé de le noter provisoirement par **n**. Mais, on sait que la **médiane** de cette série statistique est **153,125 DH**.

1. Exprimer la moyenne arithmétique de cette distribution en fonction de **n**.
2. Exprimer la médiane de cette série statistique en fonction de **n**, sachant que la valeur **N/2** n'a pas été trouvée exactement parmi les effectifs cumulés croissants.

Déterminer **n** et puis **N** le nombre d'employés de cette entreprise.

3. Retrouver la valeur numérique de la moyenne arithmétique en remplaçant la valeur de **n** trouvée dans l'expression de la moyenne exprimée en 1°.
4. Calculer la médiale de cette série statistique.
5. Représenter la courbe de concentration (ou courbe de Lorenz).

6. Calculer l'indice de concentration (ou indice de GINI) et conclure.

Exercice n° 4:

Le tableau ci-dessous donne la répartition de 200 exploitations agricoles, dans un certain pays, selon la SAU (surface agricole utilisée) exprimée en hectares :

SAU (en ha)	Fréquences (en %)
de 0 à 10	22
de 10 à 30	28
de 30 à 50	27
de 50 à 100	23

1) Déterminer le pourcentage des exploitations agricoles qui ont une SAU supérieure à 30 ha.
2) Déterminer la SAU la plus fréquente.
3) Donner la valeur et l'interprétation de la médiane (Mé).
4) Calculer le troisième quartile(Q_3) et le cinquième décile (D_5).
5) Donner l'interprétation et la valeur de la médiale (Ml).
6) Calculer la différence ΔM=Ml-Mé. Calculer l'indice de concentration.

   Interpréter le résultat.
7) Construire la courbe de Lorenz.
8) Déterminer l'indice de concentration de Gini. Conclure.

Chapitre 6 :

Ajustement et corrélation linéaire

I.- Introduction:

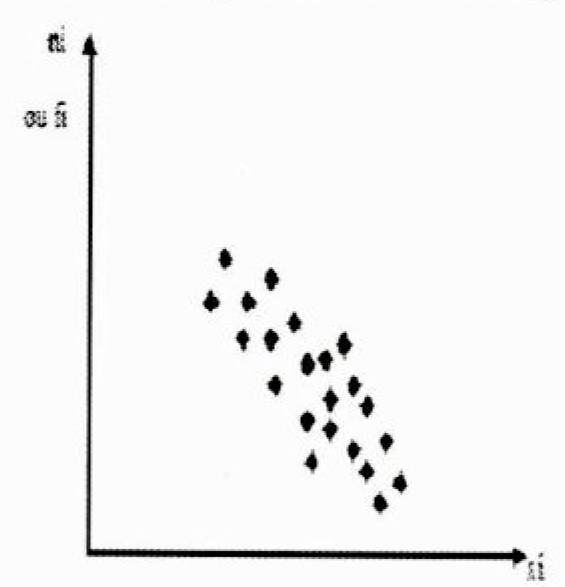
Dans ce chapitre, les variables statistiques sont supposées « discrètes ». Pour passer aux variables continues, il suffit de remplacer les classes par leurs centres.

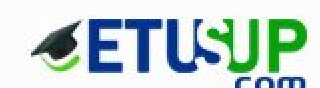
II.- Notion d'ajustement :

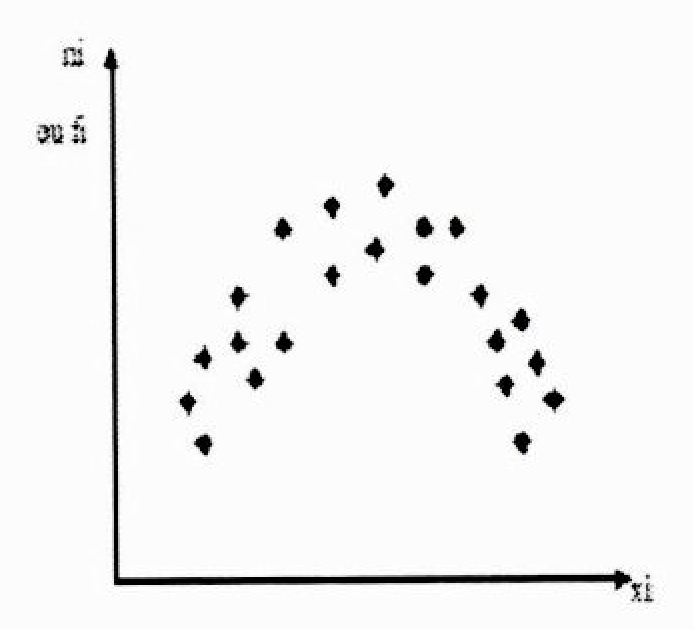
Dans les chapitres précédents, nous avons vu que les tableaux statistiques ont au moins deux colonnes :

- Une colonne pour les valeurs de la variable xi
- Une colonne pour les effectifs ni ou les fréquences fi.

Nous avons aussi procédé à des représentations graphiques.







Parfois, dans ces représentations graphiques, les points représentés semblent se repartir suivant une configuration assez régulière (une droite ou une parabole, ou une courbe).

L'ajustement consiste en substituer aux effectifs ou fréquences effectivement observés des effectifs ou fréquences calculés à l'aide de procédés que nous allons envisager.

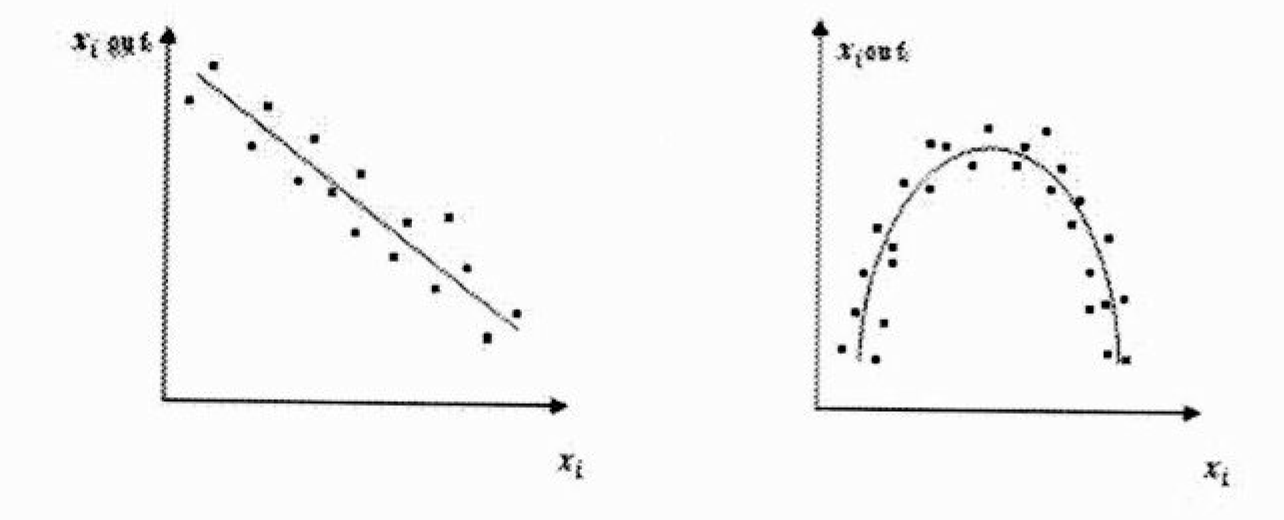
On conçoit, en examinant ces représentations graphiques, que la première étape d'une opération d'ajustement sera la recherche de la forme générale de la courbe d'ajustement, la seconde étape étant la détermination de l'équation de la courbe d'ajustement, équation telle que :

ni ou $f_i = f(x)$ avec f une fonction

1- Ajustement graphique :

Lorsqu'on a, dans une représentation graphique, un nuage de points, un première ajustement conduit à tracer une courbe simple régulière qui compense à peu près les écarts positifs ou négatifs c'est-à-dire qui laisse à peu près le même nombre de points de part et d'autre.

Ainsi les deux figures précédentes conduiraient à des ajustements manuels qui se présenteraient approximativement comme suit :



Dans le cas où les points semblent se répartir de façon linéaire, plusieurs droites peuvent-être candidates à ajuster ce nuage de points.

Mais, ces droites sont parallèles régulièrement espacées de part et d'autre d'une droite centrale. La droite d'ajustement est la droite centrale de l'ensemble des droites telle que les points à

ajuster se répartissent équitablement au dessus et au dessous de cette droite, et à l'intérieur du réseau des parallèles.

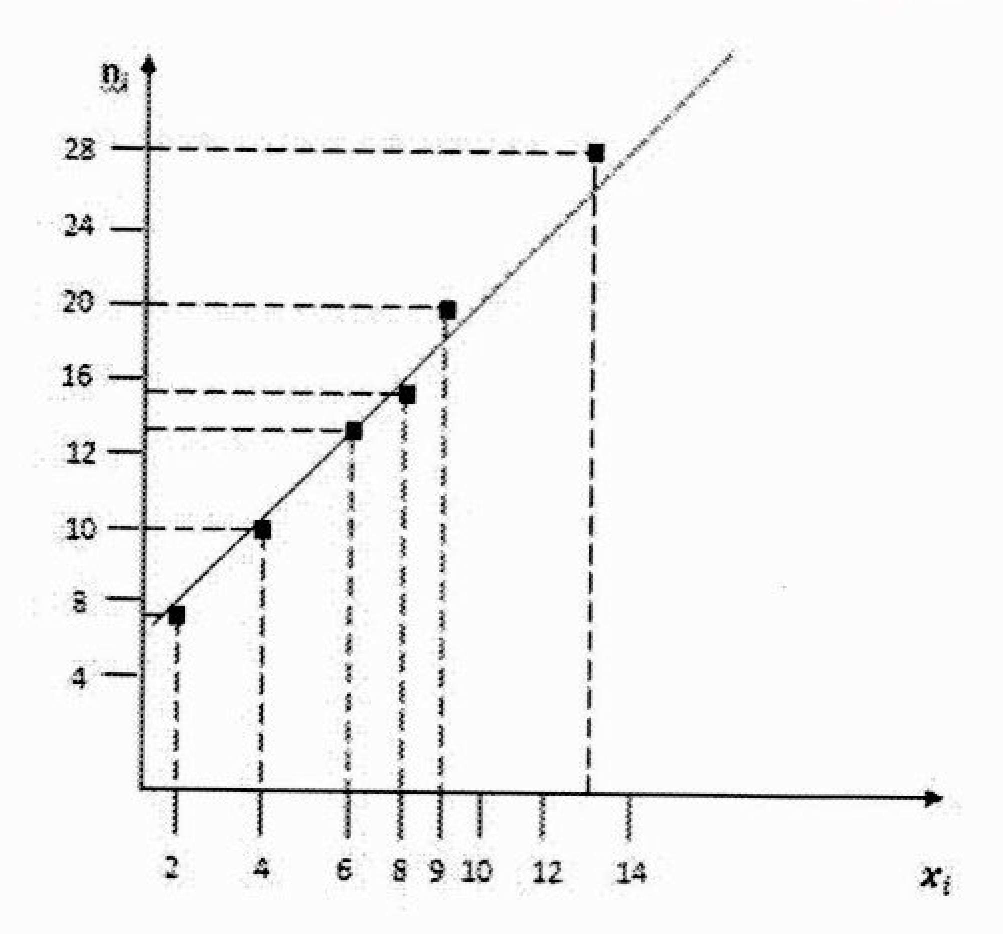
Cet ajustements manuel, bien que sommaire, donne en pratique de bons résultats surtout si la courbe d'ajustement cherchée est une courbe simple, une droite par exemple.

Exemples d'ajustements graphiques linéaires :

On dispose de tableau suivant :

x_i	n_i
2	7
4	10
6	13
8	15
9	20
13	28

La représentation graphique de cette série statistique est faite de six points.



Un ajustement linéaire est le plus approprié.

Une droite d'ajustement a été construite manuellement et figure sur la représentation. On détermine facilement l'équation $y=ax+b$ de cette droite en écrivant qu'elle passe par les points, $M(4, 10)$ et $N(9, 20)$

On peut écrire

$$\begin{cases} 10 = 4a + b \\ 20 = 9a + b \end{cases} \Rightarrow 5a = 10 \Rightarrow a = 2$$

et b=10 – 8=2

Soit une droite d'équation $y=2x+2$

2- Méthode des points médians :

Exemple :

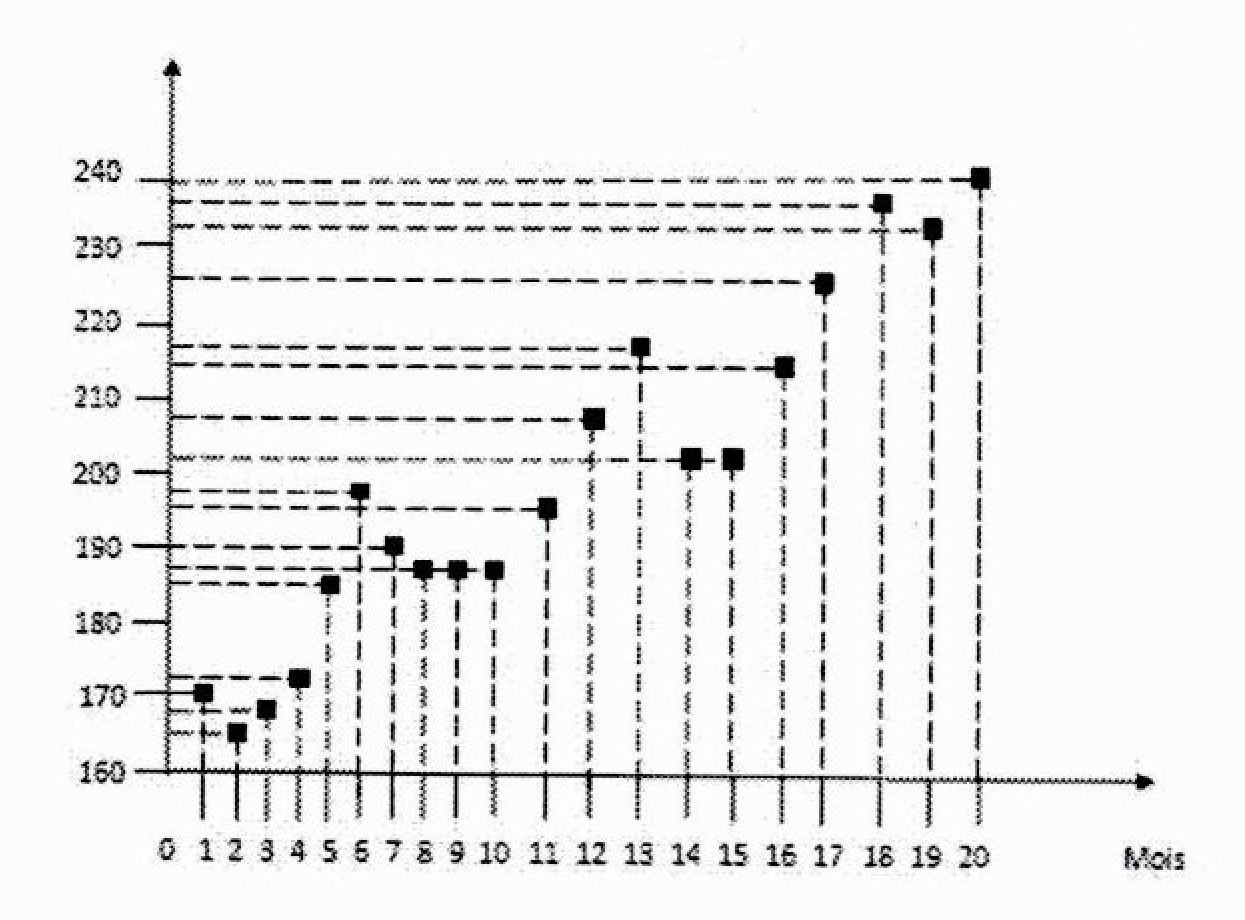
Un phénomène économique a fait l'objet d'une mesure mensuelle au cours de 20 mois consécutifs.

Représentons graphiquement la série statistique en question et ajustons par la méthode des points médians les points obtenus.

Mois	Mesures	Mois	Mesures	Mois	Mesures	Mois	Mesures
1	17	6	198	11	196	16	216
2	164	7	190	12	208	17	226
3	168	8	188	13	218	18	236
4	172	9	188	14	202	19	232
5	186	10	188	15	202	20	240

On représente d'abord les 20 points de coordonnées respectives :

(1et 170) ; (2 et 164) ; … ;(20 , 240).



III.- Notion de corrélation :

Parfois, on peut se trouver en face d'une population telle qu'on puisse étudier deux caractères différents sur une même unité statistique :

Variable 1 : X	Variable 2 : Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

n étant le nombre d'unité statistiques observées.

Une unité présente à la fois la mesure x_i de X et la mesure y_i de Y, avec $i=1, \ldots, n$

Sur le tableau précédent, on pourra s'attacher à l'étude des variations simultanées (croissance ou décroissance) des deux caractères X et Y.

Trois situations peuvent se présenter :

1) Aucun lien entre X et Y.
2) X et Y sont liés fonctionnellement.
3) Sans être liés fonctionnellement, X et Y sont en dépendance, plus ou moins marquée.

On dira alors que X et Y sont en corrélation (positive ou négative).

Exemple :

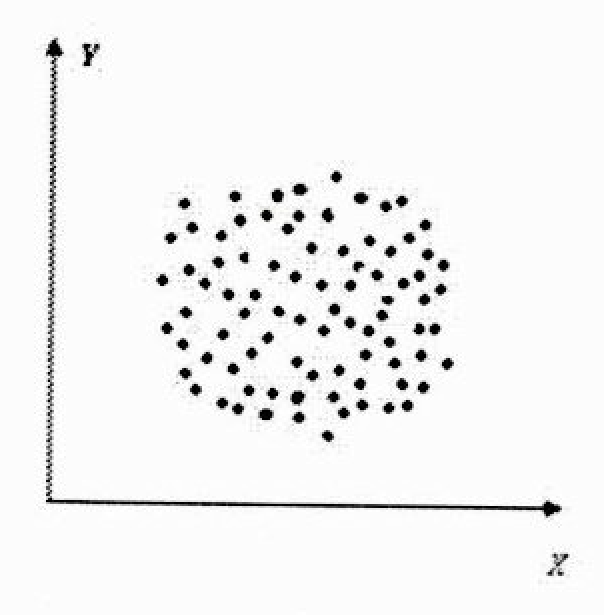
- Le poids et la taille des enfants d'une école sont deux variables en corrélation positive.

1- Mise en évidence graphique de l'existence d'une corrélation entre deux variables :

Sur un repère, on représente les points de coordonnées tirées du tableau précédent :

$$(x_1, y_1) ; (x_2, y_2) ; \ldots ; (x_i, y_i) ; \ldots ; (x_n, y_n).$$

Les points formeront un nuage de points:

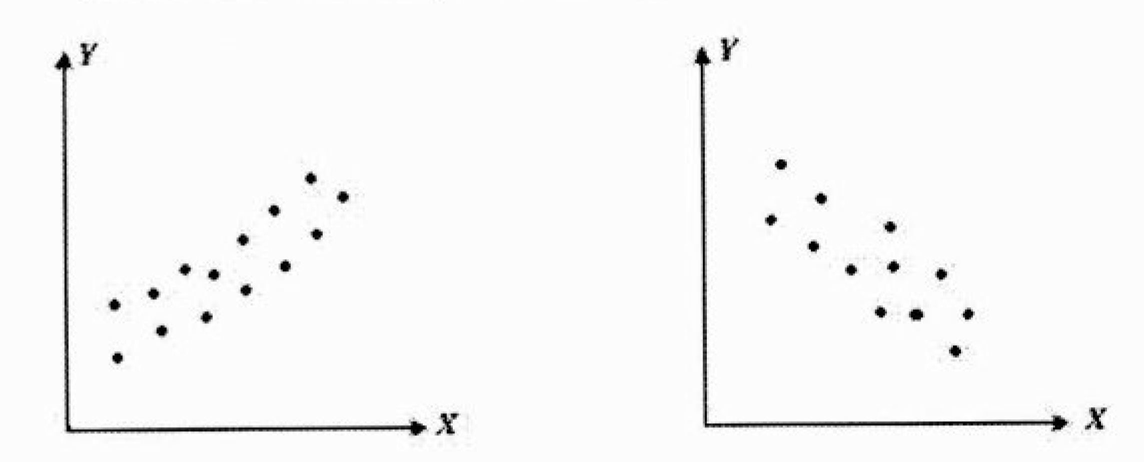


Ce nuage peut se présenter de divers façons :

a) Les points se dispersent au hasard dans le plan :

Comme dans la figure précédente. Alors X et Y peuvent être indépendantes l'une de l'autre (car le nuage ne présente pas une influence claire d'une variable sur l'autre).

b) **Les points peuvent se regrouper en un nuage de forme assez allongée, assez aplatie :**



A une valeur fixée de l'une des deux variables correspondent des points qui présentent, pour l'autre variable, des valeurs assez rapprochées.

Les deux variables semblent alors être liées, c'est-à-dire corrélées soit positivement (croissent ensembles), soit négativement (décroissement ensembles).

On peut alors dire que X et Y sont corrélées ou qu'il existe une corrélation entre X et Y.

2- Mesure de la corrélation. Coefficient de corrélation linéaire :

La correlation est mésurée par un coefficient appelé le coefficient de corrélation linéaire. Les formules de coefficient de corrélation linéaire sont :

$$r = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{X})^2 (y_i - \overline{Y})^2}}$$

$$r = \frac{\frac{1}{N}\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{\frac{1}{N}\sqrt{\sum_i (x_i - \overline{X})^2 (y_i - \overline{Y})^2}}$$

$$r = \frac{Cov(X,Y)}{\sqrt{\frac{1}{N}\sum_i (x_i - \overline{X})^2 \frac{1}{N}\sum_i (y_i - \overline{Y})^2}}$$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{Cov(X,Y)}{\sigma(X) \cdot \sigma(Y)}$$

Donc, d'après ceci, on voit que $-1 \leq r \leq 1$

Si $r = 1$, alors on a une forte corrélation positive.

Si $r = -1$, alors on a une forte corrélation négative.

Si $r = 0$, alors on n'a pas de corrélation.

Exemple de calcul du coefficient de corrélation linéaire entre deux variables X et Y :

x_i	y_i	$x_i - \overline{X}$	$y_i - \overline{Y}$	$(x_i - \overline{X})^2$	$(y_i - \overline{Y})^2$	$(x_i - \overline{X})(y_i - \overline{Y})$
16	20	-10.1	-10.4	102.01	108.16	+105.04
18	24	-8.1	-6.4	65.61	40.96	+51.84
23	28	-3.1	-2.4	9.61	5.76	7.44
24	22	-2.1	-8.4	4.41	70.56	17.64
28	32	+1.9	+1.6	3.61	2.56	3.04
29	28	+2.9	-2.4	8.41	5.76	-6.96
26	32	-0.1	-1.6	0.01	2.56	-0.16
31	36	+4.9	+5.6	24.01	31.36	+27.44
32	41	+5.9	+10.6	34.81	112.56	+62.54
34	41	+7.9	+10.6	62.41	112.36	83.74
261	304	0	0	314.90	492.40	+351.60

Moyenne arithmétique de X : $\overline{X} = \frac{261}{10} = 26,1$

Moyenne arithmétique de Y : $\overline{Y} = \frac{304}{10} = 30,4$

Coefficient de corrélation linéaire :

$$r = \frac{\sum_i (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \overline{X})^2 (y_i - \overline{Y})^2}} = \frac{351,60}{\sqrt{314,90 \times 492,40}}$$

$$= \frac{351,60}{\sqrt{155056,76}} = \frac{351,60}{393,77} = +0,89$$

Donc on a une corrélation positive, comme l'indique déjà la représentation graphique et assez serrée, le coefficient r ayant une valeur absolue voisine de 1.

3- Droite de régression linéaire: Méthode de moindres carrées:

a- Méthode des moindres carrées :

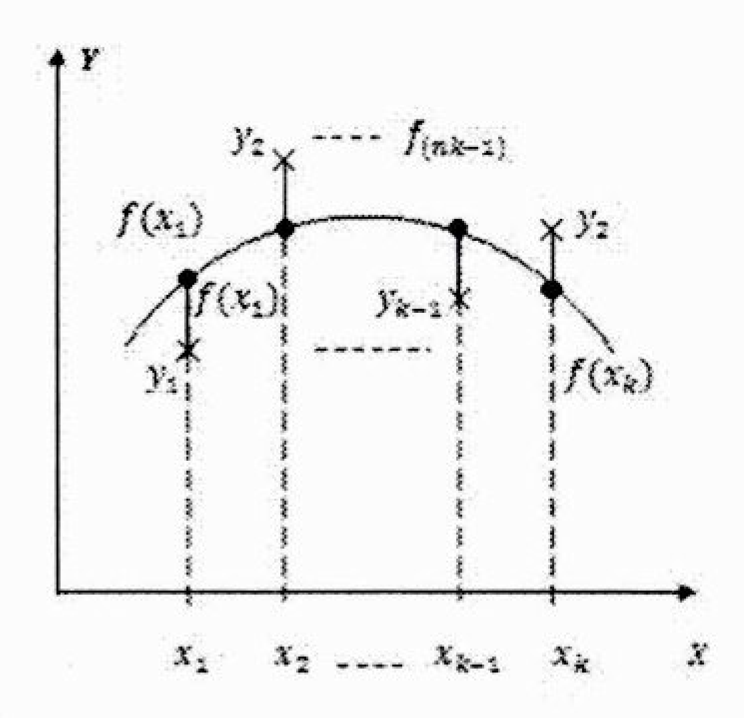
Le problème à résoudre est le suivant : il faut déterminer les différents paramètres d'une fonction on : $y = f(x)$ qui ajuste à la façon la plus satisfaisante les observations faites, la fonction retenue devant cependant conduire à une courbe d'ajustement :

x_i	y_i
x_1	y_1
$\vdots$	$\vdots$
x_k	y_k

} Observations

Supposons $y = f(x)$ est l'ajustement envisagé (on a sa forme, son type, mais pas ses paramètres!)

Graphiquement:



Calculons :

$$[y_1 - f(x_1)]^2 ; [y_2 - f(x_2)]^2 ; \cdots ; [y_{k-1} - f(x_{k-1})]^2 ; [y_k - f(x_k)]^2$$

On trouve différentes valeurs de paramètres à déterminer qui conduisent à différentes fonctions f telle que $y = f(x)$

Alors, nous cherchons les paramètres qui donnent la fonction *f* qui rend la plus faible possible la somme des carrées $[y_i - f(x_i)]^2$ (méthode des moindres carrés).

Conclusion : On cherche la fonction f telle que $y = f(x)$ qui permettra de calculer les f (x_i) tels que $\sum_i [\underset{\substack{\uparrow \\ \text{observés}}}{y_i} - \underset{\substack{\uparrow \\ \text{calculée}}}{f(x_i)}]^2$ soit minimum

On dispose d'une distribution à deux variables statistiques X et Y pour laquelle la représentation graphique a montré que Y peut-être ajusté par rapport à X par une droite d'équation $y=f(x)=ax+b$. Alors, par la méthode des moindres carrées, il faut que :

$$\min \sum_i [y_i - f(x_i)]^2 \qquad \text{c'est-à- dire} \qquad \min \sum_i [y_i - ax_i - b]^2$$

Le problème consiste donc à déterminer les paramètres a et b.

Rappelons que les valeurs numériques de x_1, … , x_k et y_1, …, y_k sont connues et lues directement sur le tableau statistique.

Nous allons calculer la dérivée par rapport à a et à b de $\sum_i [y_i - ax_i - b]^2$, puis égaler à zéro les dérivées obtenues :

$$\begin{cases} \dfrac{d\left(\sum_{i=1}^{k}[y_i - ax_i - b]^2\right)}{da} = 0 \\ \dfrac{d\left(\sum_{i=1}^{k}[y_i - ax_i - b]^2\right)}{db} = 0 \end{cases} \Rightarrow \begin{cases} 2\sum_{i=1}^{k}(y_i - ax_i - b)\times(-x_i) = 0 \\ -2\sum_{i=1}^{k}(y_i - ax_i - b) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} -2\sum_{i=1}^{k}\left(x_i y_i - ax_i^2 - bx_i\right) = 0 \\ -2\sum_{i=1}^{k}(y_i - ax_i - b) = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{k} x_i y_i - a\sum_{i=1}^{k} x_i^2 - b\sum_{i=1}^{k} x_i = 0 \\ \sum_{i=1}^{k} y_i - a\sum_{i=1}^{k} x_i - bk = 0 \end{cases}$$

$$\Rightarrow \begin{cases} +a\sum_{i=1}^{k} x_i^2 + b\sum_{i=1}^{k} x_i = +\sum_{i=1}^{k} x_i y_i & (1) \\ +a\sum_{i=1}^{k} x_i + bk = +\sum_{i=1}^{k} y_i & (2) \end{cases}$$

$$(2) \Rightarrow b = \frac{\sum_{i=1}^{k} y_i - a\sum_{i=1}^{k} x_i}{k} = \overline{y} - a\overline{x}$$

On remplace dans (1) $\Rightarrow a\sum_{i=1}^{k} x_i^2 + (\overline{y} - a\overline{x})\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} x_i y_i$

$$\Rightarrow a\left(\sum_{i=1}^{k} x_i^2 - \overline{x}\sum_{i=1}^{k} x_i\right) = \sum_{i=1}^{k} x_i y_i - \overline{y}\sum_{i=1}^{k} x_i$$

$$\Rightarrow a = \frac{\sum_{i=1}^{k} x_i y_i - \overline{y}\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} x_i^2 - \overline{x}\sum_{i=1}^{k} x_i} = \frac{\frac{1}{k}\left[\sum_{i=1}^{k} x_i y_i - \overline{y}\sum_{i=1}^{k} x_i\right]}{\frac{1}{k}\left[\sum_{i=1}^{k} x_i^2 - \overline{x}\sum_{i=1}^{k} x_i\right]}$$

$$\Rightarrow a = \frac{Cov(X,Y)}{Var(X)}$$

<u>Finalement :</u>

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var(X)} \\ b = \overline{y} - a\overline{x} \end{cases}$$

Remarque : On a une autre formule :

$$a = \frac{\sum_{i=1}^{k}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{k}(x_i - \overline{x})^2}$$

Lorsque deux variables sont en corrélation linéaire, alors on a : $y = ax + b$ est l'ajustement de y à partir de x. Cette droite s'appelle «droite de régression», on droite d'«estimation de y à partir de x ».

Le procédé de détermination de l'équation de la droite de régression est généralement fondé sur le principe des moindres carrés.

X = variable cause ; Y = variable conséquence

D'après la méthode de moindres carrées :

$$a = \frac{\sum_{i=1}^{k}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{k}(x_i - \overline{x})^2} \quad \text{et} \quad b = \overline{y} - a\overline{x}$$

Exemple :

$$a = \frac{351{,}60}{314{,}9} = 1{,}117\ ; \qquad b = 30{,}4 - (1{,}117 \times 26{,}1) = 1{,}25$$

Equation de régression : $y = 1{,}117x + 1{,}25$

Par analogie, on pourra chercher l'équation de la droite de régression de X par rapport à Y, équation de la forme $x = a'y + b'$

avec $$a' = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{k}(y_i - \bar{y})^2} \quad \text{et} \quad b' = \bar{x} - a'\bar{y}$$

Pour le même exemple on a :

$$a' = \frac{351,60}{492,40} = 0,714 \; ; \qquad b' = 26,1 - (0,714 \times 30,4) \approx 4,4$$

Equation de régression de X par rapport à Y est :

$$\boxed{x = 0,714\, y + 4,4}$$

Remarque :

On a:

$$a = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{k}(x_i - \bar{x})^2} = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2}$$

$$= \frac{Cov(X,Y)}{Var(X)}$$

$$\frac{r}{a} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \cdot \frac{\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})(y_i - \bar{y})}$$

$$= \frac{\sum(x_i - \bar{x})^2}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{Var(X)}{\sigma(X)\sigma(Y)}$$

$$\frac{r}{a} = \frac{\sigma(X)}{\sigma(Y)} \quad \text{d'où} \quad r = a\frac{\sigma(X)}{\sigma(Y)}$$

De même $$r = a'\frac{\sigma(Y)}{\sigma(X)}$$

Exemple : (l'ajustement linaire par la droite des moindres carrées)

Soit le tableau suivant (déjà vu dans l'ajustement graphique linéaire)

x_i	y_i	$x_i y_i$	x_i^2	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2	7	14	4	-5	-8.5	42.5	25
4	10	40	16	-3	-5.5	16.5	9
6	13	78	36	-1	-2.5	2.5	1
8	15	120	64	+1	-0.5	-0.5	1
9	20	180	81	+2	+4.5	9	4
13	28	364	169	+6	12.5	75	36
42	93	796	370	0	0	145	76

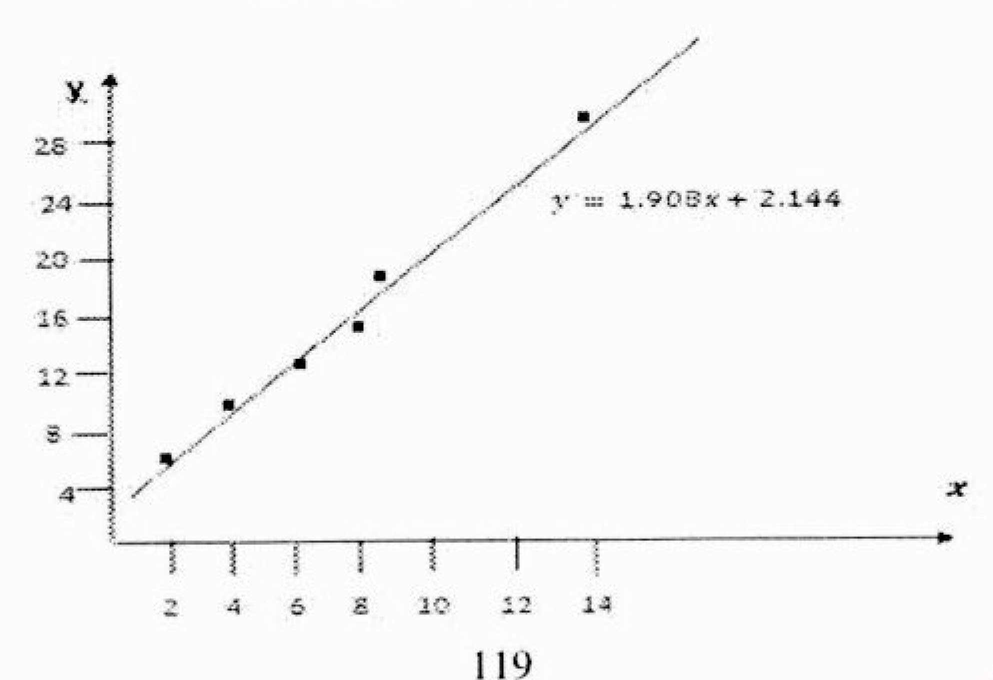
$$\bar{x} = \frac{42}{6} = 7 \qquad\qquad \bar{y} = \frac{93}{6} = 15.5$$

$$a = \frac{\sum_{i=1}^{6} x_i y_i - \bar{y}\sum_{i=1}^{6} x_i}{\sum_{i=1}^{6} x_i^2 - \bar{x}\sum_{i=1}^{6} x_i} = \frac{796-(15{,}5\times 42)}{370-(7\times 42)} = \frac{796-651}{370-294} = \frac{145}{76} \approx 1{,}908$$

$$b = \bar{y} - a\bar{x} = 15{,}5 - (1{,}908\times 7) = 15{,}5 - 13{,}356 = 2{,}144$$

Donc, l'équation de la droite d'ajustement sera:

$$\boxed{y = 1{,}908x + 2{,}144}$$



Calcul du coefficient a en utilisant la formule :

$$a = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i+\bar{x})^2} = \frac{145}{76} = 1.908$$

$$b = 2.144$$

Remarque :

Nous avons écrit dans la recherche du paramètre b

$$\sum_{i=1}^{k}(y_i - ax_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^{k}[y_i - f(x_i)] \text{ avec } f(x) = ax_i + b$$

Vérifions ce résultat sur le calcul numérique qui vient d'être fait, et en prenant $f(x) = 1.908x + 2.144$:

x_i	y_i	$f(x_i) = 1.908x + 2.144$	$y_i - f(x_i)$
2	7	5,960	+ 1,04
4	10	9,776	+ 0,224
6	13	13,592	- 0,592
8	15	17,408	- 2,408
9	20	19,316	- 0,684
13	28	26,948	+ 1,052
			0,000

Ainsi la droite des moindres carrés:

- Minimise $\Sigma(y_i - f(x_i))^2$

- Passe par les points de coordonnées $\overline{x} = \frac{\Sigma x_i}{N}$ et $\overline{y} = \frac{\Sigma y_i}{N}$.

- Rend nulle la somme algébrique des différences entre ordonnées réekkement observées, et ordonnées calculées à partir de la droite d'ajustement.

Exercices

Exercice n°1:

Une distiribution statistique se présente fe la façon suivante :

x_i	y_i
2	21
3	22
6	15
7	14
9	10
10	8
11	4
12	2

1) Représenter la graphiquement.

2) Deux statisticiens différents ont abouti aux droites d'ajustement suivantes :

- le premièr : y = -2x + 27,2
- le deuxième : y = -1,9x + 26

En se fondant sur le principe des moindres carrés apprécier lequel des deux statisticiens a réalisé le meilleur ajustement. On calculera pour les deux fonctions obtenues le total ($y_i -\ y_i$ ajustés)^2 est on comparera les résultats obtenus.

3) Effectuer ensuite l'ajustement par la méthode des moindres carrés.

4) Vérifier, à partir de l'équation de la droite des moindres carrés que le total (y_i − y_i ajustés)² est égal à zéro.

5) Vérifier, à partir de cette même équation de la droite des moindres carrés que le total (y_i − y_i ajustés)² est plus faible que chacun des deux totaux calculés en question 2).

Exercice n°2 :

Une entreprise commerciale consacre une certaine somme à des opérations publicitaires au début de chaque mois.

Dans le tableau ci-dessous sont récapitulés, mensuellement, pour 2008.

- Les sommes consacrées à ces opérations.
- Les montants des ventes correspondantes.

Mois	Dépenses de publicités x_i (en millions de dirhams)	Ventes y_i (en millions de dirhams)
Janvier	2,4	38
Février	3	42
Mars	3	42
April	2,5	39
Mai	3.2	40
Juin	3.5	45
Juillet	2	35
Août	1.8	24
Septembre	3	38
Octobre	3,2	40
Novemnre	3,8	44
Décembre	4.6	53

On demande de:

1) Calculer le coéfficient de corrélation linéaire existant entre les deux variable x et y.

2) Déterminer l'équation de la droite d'éstimation de y à partir de x.

Exercice n°3 :

Un tableau statistique se présente comme suit :

x_i	y_i
-5	33
-1	25
3	17
10	3
13	-3

1) Calculer le coéfficient de corrélation linéaire entre les variables x et y.

2) Déterminer l'équation de la fonction de régression permettant d'éstimer y à partir de la connaissance de x.

3) Représenter graphiquement le nuage des points et de la fonction de régression.

Chapitre 7 :

Séries Chronologiques

I.- Introduction:

La prévision est un élément clé pour la prise de décision. Donc, prévoir l'évolution de l'environnement économique à travers tous les facteurs qui peuvent exercer une influence sur le résultat, est une étape primordiale qui précède la prise de décision. Ce qui suppose une précision de la situation économique une fois que les choix auront été effectués, c'est enfin contrôler, c'est-à-dire comparer les résultats obtenus avec ceux qui étaient prévus.

Le plus souvent, les données utilisées dans le domaine de la prévision proviennent des séries chronologiques. C'est donc par l'étude de ce type de données que convient d'aborder l'analyse des méthodes de prévision.

Définition:

Une série chronologique ou temporelle est une suite d'observations d'un phénomène au cours du temps.

Exemple:

La statistique des ventes mensuelles d'un certain produit constitue une série chronologique, de même que l'évolution annuelle du niveau général de prix ou encore du taux d'équipement des ménages en machines à laver.

L'état du phénomène est repéré à intervalles réguliers selon une périodicité qui peut être très différente d'un problème à un autre qui dépend des possibilités d'enregistrement, mais aussi de préoccupations opérationnelles.

Toute entreprise peut avoir facilement accès à un grand nombre des séries intéressantes, soit d'origine interne (statistiques de production, de commandes, de livraisons ou encore de trésorerie), soit d'origine externe (statistiques de syndicats professionnels, de la direction nationale des statistiques). Mais ces données ne sont généralement pas utilisables dans leur état.

La formalisation puis le traitement des séries chronologiques sont deux étapes préalables à leur utilisation en matière de prévision ou de contrôle.

II.- Formalisation des séries chronologiques :

Toute série chronologique apparaît comme le résultat de plusieurs mouvements élémentaires qui se superposent.

Un modèle de série chronologique devra mettre en évidence quelles sont ces composantes et quelles sont ses règles de composition.

1.- Composantes des séries chronologiques:

L'évolution d'un phénomène au cours du temps est généralement déterminée par le mouvement de quatre composantes élémentaires :

- La tendance Tt
- La conjoncture Ct
- La saison St
- Le résidu Rt

a.-La tendance (Trend): elle représente l'évolution moyenne à long terme du phénomène étudié (on l'ajuste par une droite en général).

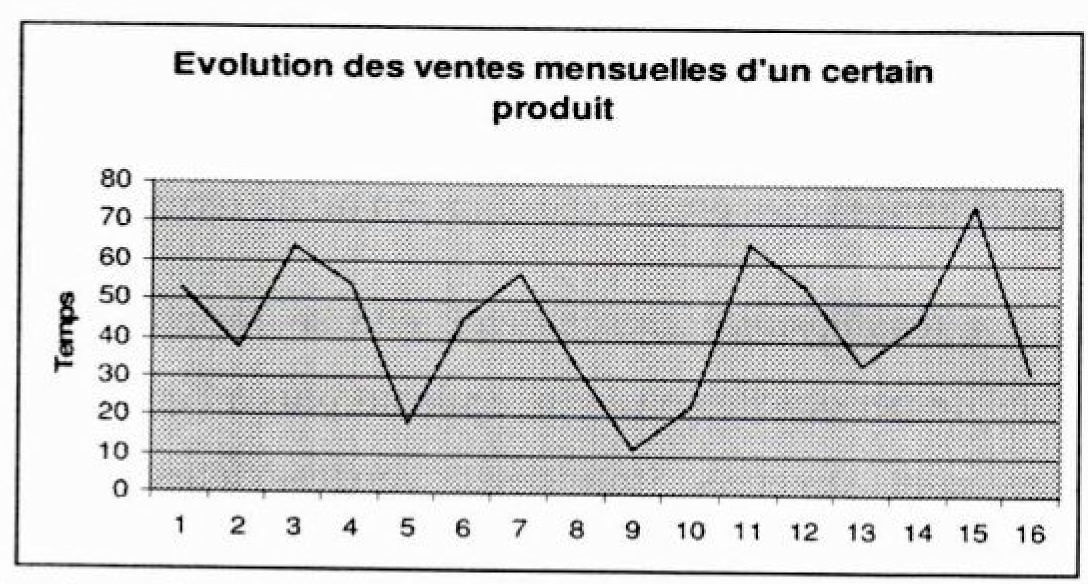
b.-La conjoncture (cycle): c'est un facteur qui joue à moyen terme et qui se manifeste par des mouvements cycliques faisant alterner, sur plusieurs années, des phases d'expansion ou de récession.

c.-La saison (composante saisonnière): elle se rapporte à des fluctuations de périodicité constante égale au plus à l'année, dont les causes sont diverses (traditions, congés, fêtes, saisons, vacances, météorologie…) (court terme).

d.-Le résidu (facteur résiduel): il recouvre un ensemble hétérogène d'éléments que l'on regroupera en deux catégories selon qu'ils introduisent des perturbations identifiables et donc éliminables ou au contraire aléatoires. C'est une composante accidentelle ou imprévisible dépendante du hasard et de l'aléatoire (Innovations technologiques imprévues, catastrophes naturelles, grève, …)

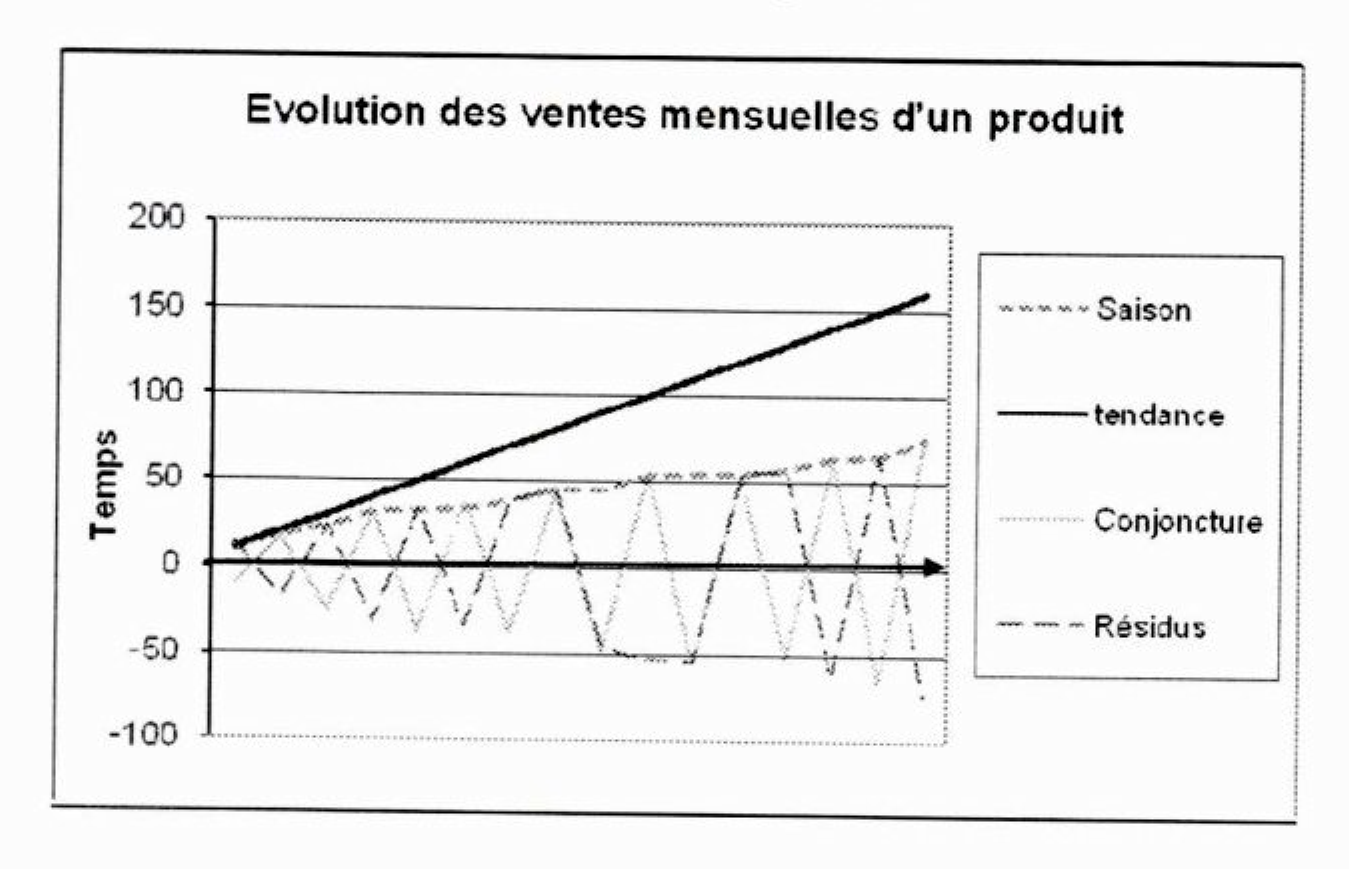
2.- L'analyse graphique des composantes:

Une série chronologique se définit comme la succession des valeurs d'une variable ordonnées par un indicateur de temps. Elle se représente généralement sous la forme d'un graphe tel que le suivant :



L'analyse des séries chronologiques considère la décomposition en quatre types de variations: la tendance, la conjoncture, la saison, et le résidu.

Le schéma suivant illustre cette décomposition :



3.- Règle de composition des séries chronologiques:

la valeur observée à la date **t** de la variable étudiée **Yt**, est fonction des valeurs prises à cette même date par les diverses composantes :

$$Yt = f\,(Tt, Ct, St, Rt)$$

Deux schémas de composition principaux sont utilisés :

- le modèle additif,
- et le modèle multiplicatif.

➢ Dans le modèle additif, la valeur **Yt** est la somme des valeurs prises par les différentes composantes:

$$Yt = Tt + Ct + St + Rt$$

où St=St+m si le cycle saisonnière porte sur **m** périodes.

Dans ce modèle,

➲ Chaque élément est mesuré dans la même unité que Yt.

➲ Si la variation saisonnière manifeste une périodicité de **m** périodes alors: $\sum_{t=i+1}^{i+m} St = 0$: dans la mesure où la tendance

représente la moyenne du phénomène, il faut que les variations saisonnières amenant **Yt** au dessus de la tendance soient compensées par des variations d'amplitude équivalente au dessous.

- De façon plus générale, et pour la même raison, si la tendance a été correctement calculée, l'effet moyen à long terme de chacun de ces éléments C, S et R doit être nul.

Remarque:

Dans le modèle additif, l'effet d'une variation de l'une des composantes sur **Yt** est indépendant de la valeur prise par les autres composantes.

- Dans le Modèle multiplicatif, **Yt** apparaît comme le produit des différentes composantes :

$$Yt = Tt \,.\, Ct \,.\, St \,.\, Rt,$$

où **Yt** est mesuré en unités (dhs, tonnes, etc....) de même que la tendance ; les autres facteurs sont intégrés dans le modèle sous la forme de coefficients multiplicateurs de type **St = (1+st),** où **st** est

le pourcentage de variation due à la saison, par exemple.

Le modèle multiplicatif peut être traduit sous forme additif si l'on utilise des logarithmes avec:

$$\text{Log } Yt = \text{Log } Tt + \text{Log } Ct + \text{Log } St + \text{Log } Rt$$

Dans la mesure où les fluctuations saisonnières doivent se compenser sur **m** périodes dans le modèle additif, il faut que

$$\sum_{t=i+1}^{i+m} Log\ St = 0,$$

ce que revient à dire, puisque $\sum Log(St) = Log\left(\prod St\right)$ que le produit des coefficients saisonniers doit être égal à 1. En fait, la condition $\prod_{t=i+1}^{i+m} St = 1$, sera souvent approchée par la condition $\sum_{t=i+1}^{i+m} St = m$.

En effet, quand « ***si*** » est petit,

on peut approcher $Log(Si) = Log(1+si)$ par si.

Dans ce cas, $\sum Log(Si)$ est équivalent à $\sum si$.

La condition $\sum Log(Si) = 0$ devient alors $\sum(1+si) = m$.

Dans le modèle multiplicatif, les composantes ne sont plus indépendantes. L'effet saisonnier sera d'autant plus important en valeur absolue, que la tendance et la conjoncture seront élevées.

À priori, aucun de ces deux modèles ne doit être déclaré supérieur à l'autre: il convient de les tester sur les données afin de repérer celui qui reproduit le mieux le phénomène considéré. Il faut noter également que des formulations mixtes peuvent être utilisées, comprenant une partie additive et une partie multiplicative.

III.- Traitement des séries chronologiques:

Traiter une série chronologique consiste à mettre en évidence ces différentes composantes à l'aide d'une procédure de décomposition donnée.

La technique la plus courante fait appel aux moyens mobiles pour dégager l'évolution à moyen terme du phénomène, c'est-à-dire la combinaison de T et de C (c'est-à-dire de minimiser les fluctuations aléatoires). Une fois cette opération réalisée, il devient possible de déterminer la valeur des coefficients saisonniers, de même que la composante résiduelle.

1.- La méthode des moyennes mobiles:

Étant donné un ensemble d'observations correspondantes à une série chronologique, la détermination des moyennes mobiles d'amplitude « h » se fait de la façon suivante:

i. On forme des groupes de « h » observations et on calcul leurs moyenne.

ii. Le premier groupe est formé par les « h » premières observations.

iii. Les groupes suivants seront formés en excluant du groupe antérieur la première observation et en incluant la postérieure (observation) dans le groupe suivant.

iv. On continue le processus jusqu'à ce qu'on ne peut plus former des groupes

Les moyennes obtenues dans le processus antérieur s'appellent moyennes mobiles d'amplitude ou d'ordre « h ».

Chaque moyenne mobile sera associée au point du milieu de l'intervalle de temps sur lequel elle a été calculée.

Nous allons voir ça pour h=3

Données		Somme des groupes de trois observations	Moyennes mobiles d'ordre 3
t1	Yt1		
t2	Yt2	Yt1+Yt2+Yt3	$\overline{Y2}$=1/3[Yt1+Yt2+Yt3]

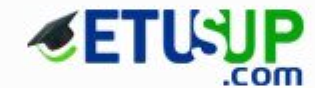
t3	Yt3	Yt2+Yt3+Yt4	$\overline{Y3}$=1/3[Yt2+Yt3+Yt4]
t4	Yt4	Yt3+Yt4+Yt5	$\overline{Y4}$=1/3[Yt3+Yt4+Yt5]
...	...	...	...
tn-2	Ytn-2	Ytn-3+Ytn-2+Ytn-1	$\overline{Y}$n-2=1/3[Ytn-3+Ytn-2+Ytn-1]
tn-1	Ytn-1	Ytn-2+Ytn-1+Ytn	$\overline{Y}$n-1=1/3[Ytn-2+Ytn-1+Ytn]
tn	Ytn		

Remarquons qu'il y a des observations pour lesquels, il n'y a pas de moyennes mobiles (pour la première et la dernière observations du tableau précédent), par.conséquent dans les calculs posterieurs des moyennes mobiles on aura moins de données que dans les "n" de départ.

La méthode des moyens mobiles repose sur le príncipe suivant: Les variations saisonnières disparaissent pratiquement lorsqu'on remplace les valeurs observées de la série par des moyennes calculées sur une période suffisamment longue (3 mois, 4 mois, 5 mois,..., ou 12 mois).

La période doit être choisie en fonction de la périodicité des résultats: s'il s'agit de variations saisonnières sur une année, la période est 12 mois, 4 trimestres ou 3 quadrimestres...

La méthode des moyennes mobiles est une méthode de lissage, c'est-à-dire qu'elle atténue les variations d'une série



chronologique. Comme telle, elle peut être employée pour lisser une série, que celle-ci présente ou non des variations saisonnières. Les moyennes mobiles que nous avons considéré sont des moyennes avec des pondérations égales pour chacune des observations (parfois on les appelle non-pondérées), il est possible aussi de considérer un ensemble de pondérations associées.

Il existe ainsi des moyennes mobiles pondérées, dans lesquelles on donne, en général, un poids plus important aux observations centrales, et moins important aux observations plus éloignées du centre.

Si le nombre des observations choisies pour les moyennes mobiles est pair, il sera nécessaire de les centraliser dans les mêmes instants de temps où se situent les observations initiales.

Le tableau suivant expose les opérations nécessaires pour h=4.

Données		Somme des groupes de 4 observations	Moyennes mobiles d'ordre 4	Moyenne mobiles centrées
t_1	Y_{t1}			
t_2	Y_{t2}			
t_3	Y_{t3}	$\sum Y_{tj}$	$\overline{Y_{2.5}}=1/4\sum Y_{tj}$	$\overline{Y_3}=1/2[\overline{Y_{2.5}}+\overline{Y_{3.5}}]$
t_4	Y_{t4}	$\sum Y_{tj}$	$\overline{Y_{3.5}}=1/4\sum Y_{tj}$	$\overline{Y_4}=1/2[\overline{Y_{3.5}}+\overline{Y_{4.5}}]$
t_5	Y_{t5}	$\sum Y_{tj}$	$\overline{Y_{4.5}}=1/4\sum Y_{tj}$	$\overline{Y_5}=1/2[\overline{Y_{4.5}}+\overline{Y_{5.5}}]$

t_6	Y_{t6}	$\sum Y_{tj}$	$Y_{5.5}=1/4\sum Y_{tj}$	$Y_6=1/2[Y_{5.5}+ Y_{6.5}]$

Centrer des moyennes mobiles d'ordre pair est équivalent à calculer des moyennes mobiles pondérées d'ordre impair.

En effet, vérifions cette affirmation sur le tableau précédent:

$$\overline{Y}_3 = \frac{1}{2}\left(\overline{Y}_{2.5} + \overline{Y}_{3.5}\right) = \frac{1}{2}\left(\frac{1}{4}\sum_{j=1}^{4} Yt_j + \frac{1}{4}\sum_{j=2}^{5} Yt_j\right)$$

$$= \frac{1}{2}\left(\frac{1}{4}\left(Yt_1 + 2Yt_2 + 2Yt_3 + 2Yt_4 + Yt_5\right)\right)$$

$$= \frac{1}{8}Yt_1 + \frac{1}{4}Yt_2 + \frac{1}{4}Yt_3 + \frac{1}{4}Yt_4 + \frac{1}{8}Yt_5$$

Comme on peut remarquer, la moyenne mobile centrée d'ordre 4 ($\overline{Y}3$) calculée pour l'instant t_3, est la moyenne mobile pondérée d'ordre 5 où les valeurs extrèmes ont un poids égal à la moitié du poids correspondant aux valeurs centrales.

De même, on le peut vérifier pour n'importe quel autre instant t_i.

Exemple:

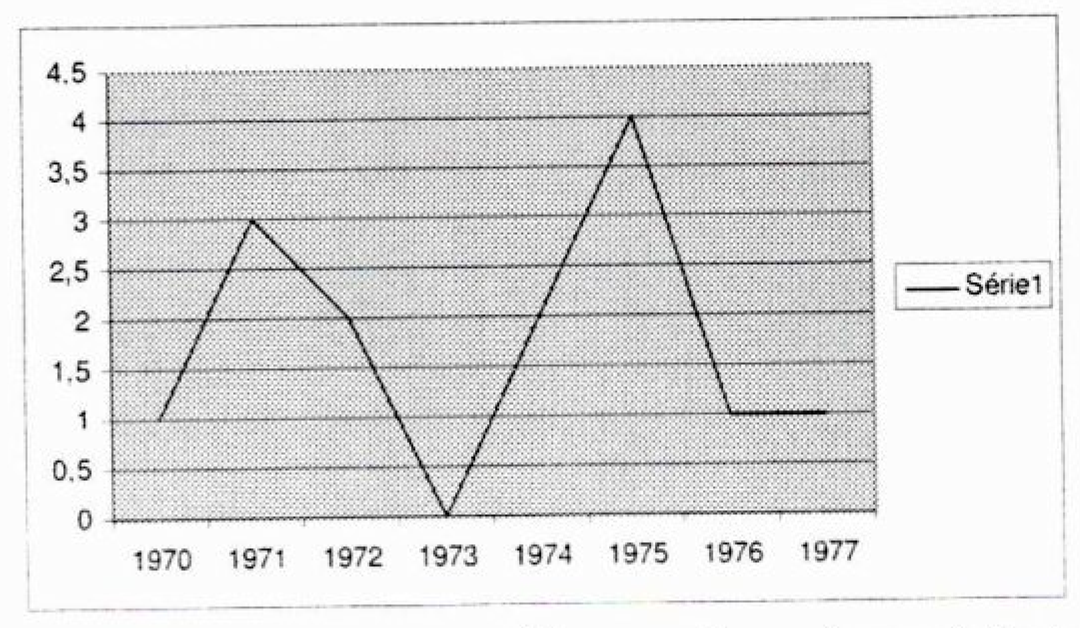
Calculer les moyennes mobiles d'amplitude 3 et d'amplitude 4 pour les données suivantes :

t	*Yt*	*h*=3		*h*=4		
1970	1	-	--			--
1971	3	6	2			

1972	2	5	1,66	6	1,5	
1973	0	4	1,33	7	1,75	1,625
1974	2	6	2	8	2	1,875
1975	4	7	2,33	7	1,75	1,875
1976	1	6	2	8	2	1,875
1977	1					

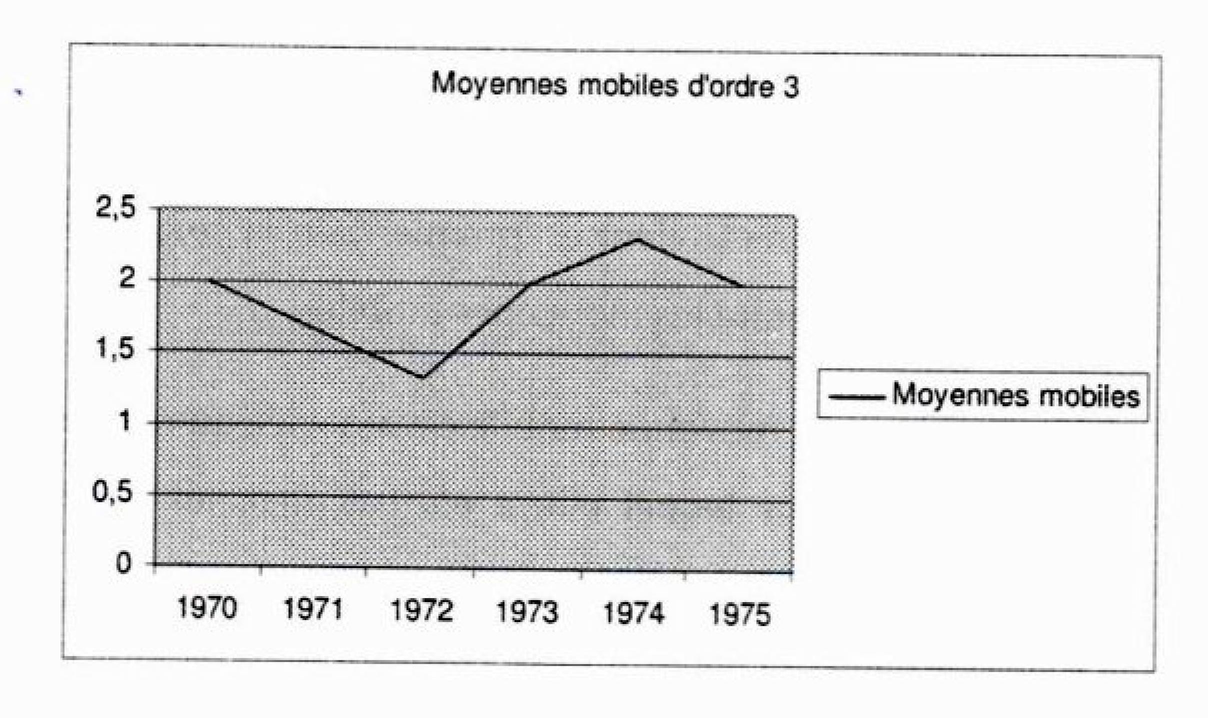
Sur les colonnes h=3 et h=4, vient en premier lieu la colonne des sommes mobiles et à la suite les moyennes mobiles.

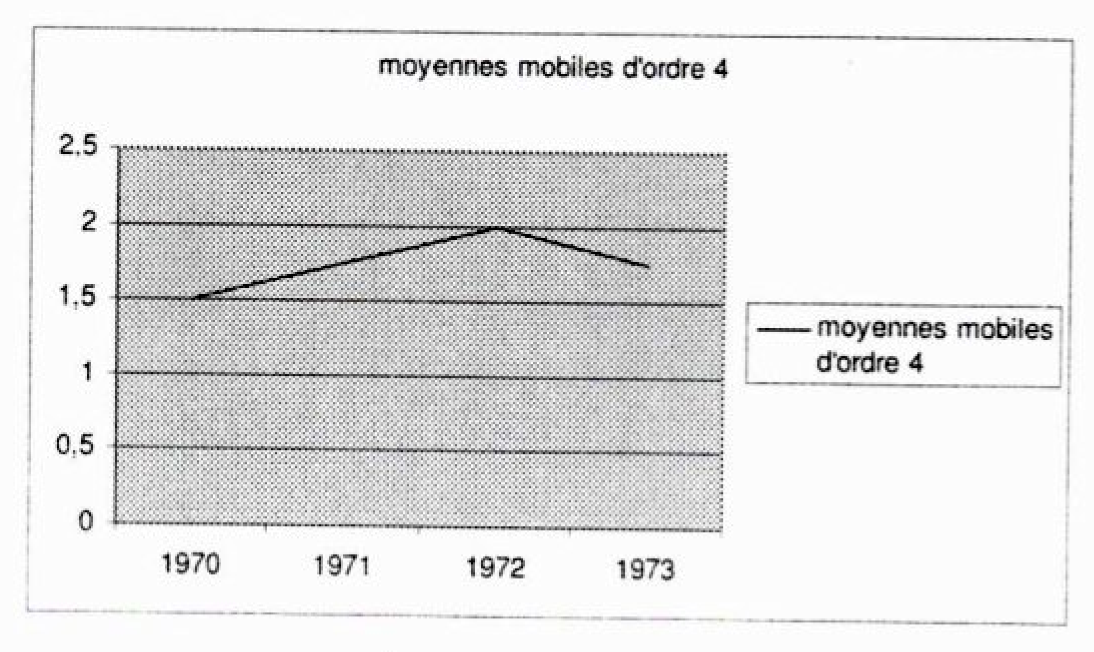
Étant donné que chaque moyenne mobile est associée au milieu de l'intervalle de temps sur lequel elle a été calculée, les moyens d'ordre trois se fixent dans le deuxième des trois années sur lesquelles elles ont été calculées.



La méthode des moyennes mobiles a deux inconvénients:

premièrement, elle ne donne pas de moyennes mobiles pour le premier et le dernier groupe de périodes de la série. Au cas où la série chronologique serait composée d'un nombre limité d'observations, les valeurs omises peuvent représenter une importante perte d'information.





Deuxièment, les moyennes mobiles "négligent" la plupart des valeurs précédentes de la série chronologique.

Ces deux inconvénients sont redressés par la méthode de lissage exponentielle

2.- Méthode de lissage exponentielle:

La méthode exponentielle de lissage d'une série chronologique est définie de la façon suivante : $X_t = \theta Yt + (1-\theta)X_{t-1}$ pour $t \geq 2$.

Où X_t est la série chronologique lissée exponentiellement à la date t.

Yt est la série chronologique à la date t

X_{t-1} est la série chronologique lissée exponentiellement à la date t-1

θ est le coefficient de lissage ; avec $0 \leq \theta \leq 1$

On commence par poser:

$Y_1 = X_1$

Ce qui donne $X_2 = \theta Y_2 + (1-\theta)X_1$

$$= \theta Y_2 + (1-\theta)Y_1$$

$$X_3 = \theta Y_3 + (1-\theta)X_2$$

$$= \theta Y_3 + (1-\theta)(\theta Y_2 + (1-\theta)Y_1)$$

$$= \theta Y_3 + (1-\theta)\theta Y_2 + (1-\theta)^2 Y_1$$

$$X_4 = \theta Y_4 + (1-\theta)X_3$$
$$= \theta Y_4 + (1-\theta)\left(\theta Y_3 + (1-\theta)\theta Y_2 + (1-\theta)^2 Y_1\right)$$
$$= \theta Y_4 + (1-\theta)\theta Y_3 + (1-\theta)^2\theta Y_2 + (1-\theta)^3 Y_1$$

En générale, on obtient:

$$X_t = \theta Y_t + (1-\theta)\theta Y_{t-1} + (1-\theta)^2\theta Y_{t-2} + \cdots + (1-\theta)^{t-1} Y_1$$

Cette dernière formule indique qu'une série chronologique « lissée » à la date t, dépend de toutes les observations antérieures de la série.

Le coefficient de lissage θ est choisi en fonction du degré de lissage demandé. Une valeur de θ tendent vers 0 (zéro) produit un degré de lissage assez important. Par contre, une valeur θ proche de 1 résulte un lissage assez limité de la série en question.

Exemple:

Pour pouvoir élaborer un système de prévision des ventes de sa filiale, une firme multinationale a relevé, pendant une période de 4 semaines, les ventes quotidiennes de la filiale en question.

Les résultats de ce suivie sont regroupés dans le tableau suivant:

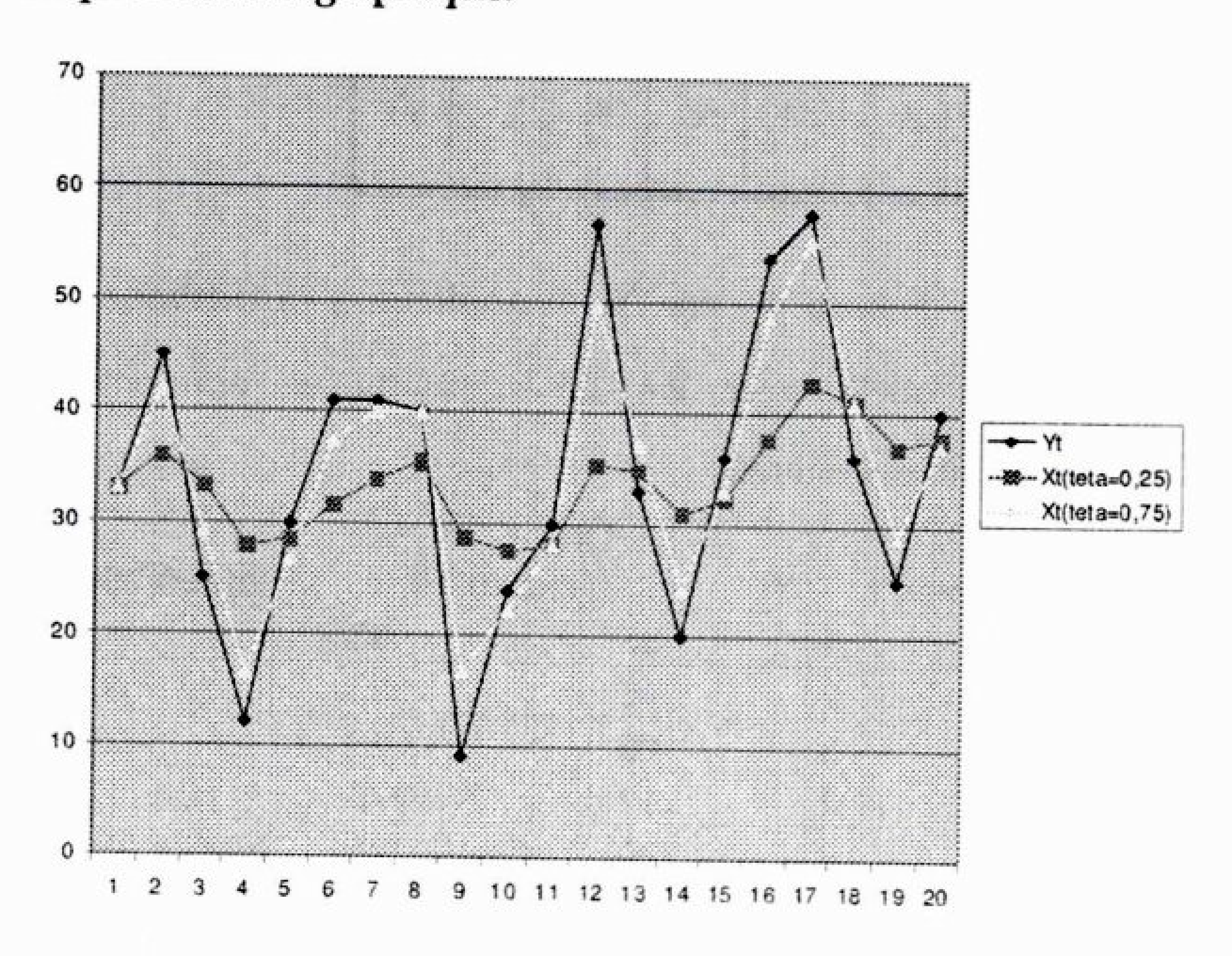
		Jours				
	i \ j	1	2	3	4	5
Semaines	1	33	45	25	12	30
	2	41	41	40	9	24
	3	30	57	33	20	36
	4	54	58	36	25	40

Transformons d'abord Y_{ij} en Yt et appliquons la méthode de lissage exponentielle avec $\theta = 0{,}25$ et $\theta = 0{,}75$ et représentons graphiquement les résultats.

	Dates t	Yt	Xt (teta=0,25)	Xt(teta=0,75)
S1	1	33	33	33
	2	45	36	42
	3	25	33,25	29,25
	4	12	27,9375	16,3125
	5	30	28,453125	26,578125
S2	6	41	31,58984375	37,39453125
	7	41	33,94238281	40,09863281
	8	40	35,45678711	40,0246582
	9	9	28,84259033	16,75616455

	10	24	27,63194275	22,18904114
S3	11	30	28,22395706	28,04726028
	12	57	35,4179678	49,76181507
	13	33	34,81347585	37,19045377
	14	20	31,11010689	24,29761344
	15	36	32,33258016	33,07440336
S4	16	54	37,74943512	48,76860084
	17	58	42,81207634	55,69215021
	18	36	41,10905726	40,92303755
	19	25	37,08179294	28,98075939
	20	40	37,81134471	37,24518985

Représentation graphique:



Remarque:

Les moyennes mobiles et la méthode exponentielle sont des méthodes relativement assez limitées de réduction ou d'élimination des fluctuations aléatoires dans le but de découvrir l'existence d'autres composantes.

3.- Détermination de la variation saisonnière:

Les fluctuations saisonnières correspondent à des variations qui peuvent avoir lieu durant une année ou même sur un intervalle de temps plus court: mois, semaine, jour. Dans le but de mesurer l'effet saisonnier, nous allons construire des indices ou coefficients saisonniers, qui ont pour objet de mesurer le degré de différence entre les saisons.

Une condition nécessaire pour l'étude de la composante saisonnière est qu'on ait une série chronologique suffisamment longue pour qu'on puisse observer l'existence de saisons.

Les indices saisonniers sont calculés de la façon suivante:

1. Extraire l'effet des fluctuations saisonnières en calculant les moyennes mobiles. Si on utilise par exemple un modèle multiplicatif de série chronologique, on aura:

$$Y_t = T_t \times C_t \times S_t \times R_t$$

   Les moyennes mobiles éliminent S_t et R_t.

2. Calculer le rapport de la série sur la moyenne mobile. On obtient:

$$\frac{Y_t}{T_t \times C_t} = S_t \times R_t$$

Le résultat obtenu est une mesure de la variation saisonnière et aléatoire.

3. Pour chaque type de saison, calculer la moyenne des rapports obtenus au 2ème point. Cette procédure extrait ou élimine la majorité de la variation saisonnière. Cette moyenne est au fait une mesure des différences saisonnières.
4. Les indices saisonniers sont les rapports moyens obtenus au 3ème point ajustés pour s'assurer que l'indice saisonnier moyen est égal à l'unité.

Exemple:

Le nombre de visiteurs d'une grande galerie d'exposition pendant les trois dernières années est consigné dans le tableau suivant:

		Trimestres			
	j / i	1	2	3	4
Années	1	10	9	12	15,5
	2	14	14	18	21
	3	22	21,5	25	30

Calculer les indices saisonniers pour chaque trimestre pour pouvoir mesurer le volume de variation saisonnière.

Dates t	Yt	M.M. d'ordre 4	M.M.C. d'ordre 4	Yt/MMC4
1	10			
2	9			
3	12	11,625	12,125	0,989690722
4	16	12,625	13,25	1,169811321
5	14	13,875	14,625	0,957264957
6	14	15,375	16,0625	0,871595331
7	18	16,75	17,75	1,014084507
8	21	18,75	19,6875	1,066666667
9	22	20,625	21,5	1,023255814
10	22	22,375	23,5	0,914893617
11	25	24,625		
12	30			

Si l'on regroupe les rapports par trimestre, nous pouvons constater les similarités pour chaque type de trimestre et les différences entre les différentes types de trimestres.

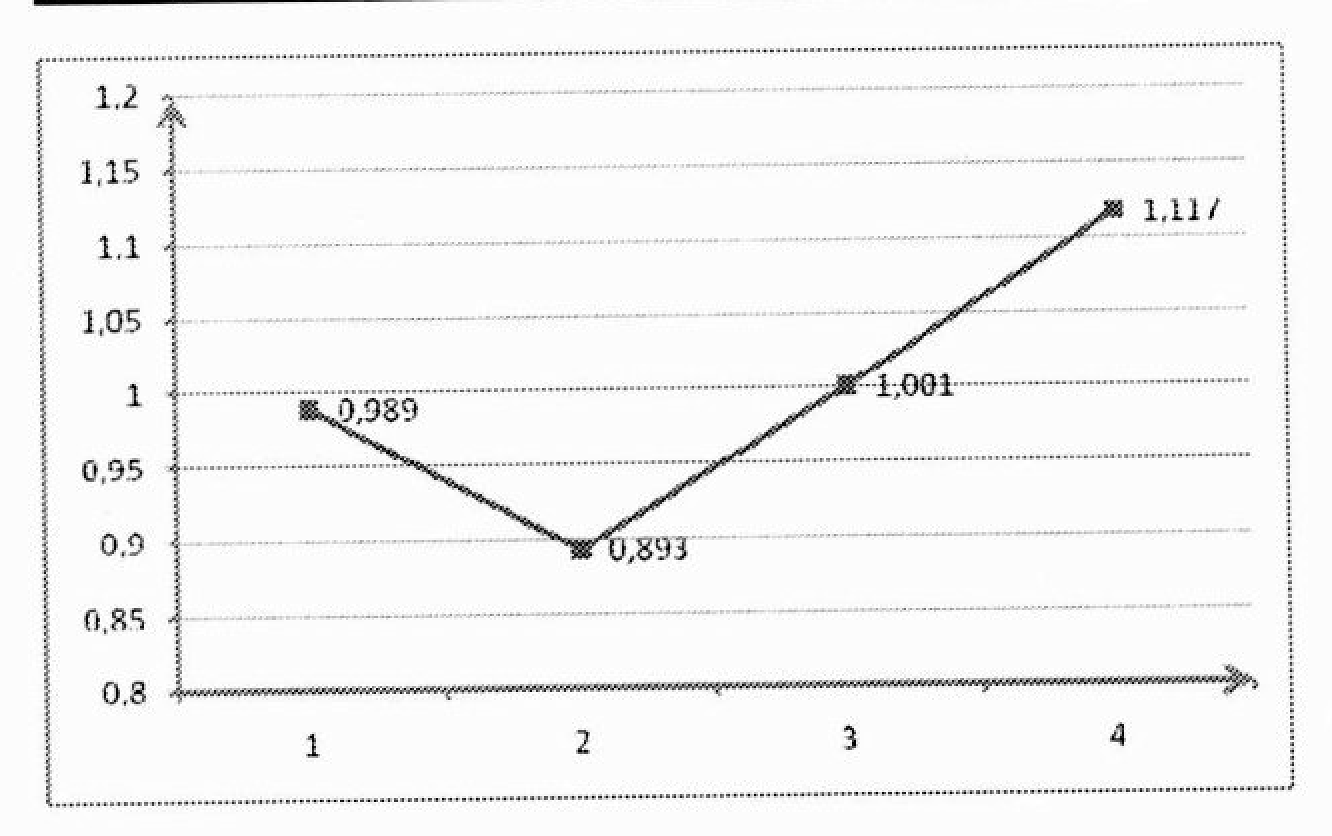
Une fois, les données sont groupées par trimestre, on calculera la moyenne des valeurs pour chaque trimestre, pour éliminer ou épurer la variation aléatoire.

La dernière étape consiste à ajuster les moyennes en divisant chacune d'elles par la somme des quatres moyennes et en multipliant par 4 (nombre de trimestres par an), on obtiendra alors les indices saisonniers.

Trimestres / Années	1	2	3	4	Total
1	—	—	0,99	1,17	
2	0,957	0,872	1,014	1,067	
3	1,023	0,915	—	—	
Moyennes	0,99	0,893	1,002	1,118	4,004
Coefficient Saisonnier (S_i)	0,989	0,893	1,001	1,117	4

Les indices saisonniers indiquent qu'en moyenne, le nombre de visiteurs de la galerie des 1er et 2ème trimestre sont inférieures à la moyenne annuelle, alors que le nombre de visiteurs pendant le 4ème trimestre dépasse la moyenne annuelle.

Representation graphique des índices saisonniers:



4.-Détermination de la variation résiduelle:

Une fois les coefficients saisonniers, S_t, déterminés, l'évolution des variations résiduelles s'effectue en prenant le rapport des combinaisons $(S_t.R_t)$ à S_t. C'est ce à quoi on a procédé dans le tableau suivant:

Trimestres / Années	1	2	3	4
1	—	—	0,989	1,047
2	0,968	0,976	1,013	0,955
3	1,034	1,025	—	—

Exercices

Exercice n°1:

On considère les ventes trimestrielles d'un produit depuis 4 ans (Ventes en milliers d'unités) :

Trimestres	1	2	3	4
Ventes première année	150	80	110	205
Ventes deuxième année	170	95	125	215
Ventes troisième année	180	105	115	240
Ventes quatrième année	195	110	150	255

1. Calculer les moyennes mobiles de longueur 4.
2. Représenter graphiquement les ventes trimestrielles et les moyennes mobiles (sur le même graphique). Qu'est ce que vous remarquez ?
3. Calculer les indices saisonniers pour chaque trimestre.
4. Déterminer la variation résiduelle.
5. Appliquer la méthode exponentielle de lissage avec θ=0,75 et représenter graphiquement les résultats.

Exercice n°2:

Pour pouvoir élaborer un système de prévision des ventes de sa filiale, une firme multinationale a relevé pendant une période de 4 ans, les ventes trimestrielles de la filiale en question.

Les résultats de suivie sont regroupés dans le tableau suivant :

		Trimestres			
		1	**2**	**3**	**4**
Années	**1**	33	45	25	12
	2	41	41	40	9
	3	30	57	33	20
	4	54	58	36	25

1. Calculer les moyennes mobiles de longueur 4.
2. Représenter graphiquement les ventes trimestrielles et les moyennes mobiles (sur le même graphique). Qu'est ce que vous remarquez ?
3. Calculer les moyennes mobiles centrées de longueur 4.
4. Calculer les indices saisonniers pour chaque trimestre.

5. Déterminer la variation résiduelle.
6. Appliquer la méthode exponentielle de lissage avec **θ=0,75** et représenter graphiquement les résultats.
7. La direction de l'entreprise estime que la tendance des ventes de sa filiale est linéaire. Analyser la tendance de ces ventes en utilisant la méthode des moindres carrée.
8. Étudier la qualité de cette représentation en calculant le coefficient de corrélation. Conclure.

Chapitre 8 :

Les nombres indices

I- Introduction et définition :

On désigne par indice la grandeur statistique avec laquelle on mesure les variations dans le temps d'une variable ou d'un ensemble de variables dépendantes. On appelle tableaux d'indices, l'ensemble des indices correspondant à plusieurs années, plusieurs localités, etc…

A l'aide des indices on pourra, par exemple, comparer le coût de la vie dans une ville, durant une certaine année, avec celui des années précédentes, ou, par exemple, la production annuelle d'acier d'un pays avec celle d'un autre pays…etc. Les indices trouvent leurs applications dans de nombreux domaines : les affaires, l'économie, l'éducation la sociologie, …etc.

C'est pourquoi de nombreux organismes gouvernementaux et privés sont engagés dans le calcul des indices. Afin de prévoir les affaires et les conjonctures économiques. Ces organismes rassemblent sans cesse des informations d'ordre général. On distingue ainsi l'indice des salaires, de la production, du chômage, des naissances, … etc.

Le plus connu, peut être le plus exploité, est celui du coût de la vie ou indice des prix à la consommation, mis au point par le « Bureau of Labor Statistics ». Au Maroc, les indices et les indicateurs économiques sont établis par la « Direction de la

Statistique » (côté public) et par le « Centre Marocain de Conjoncture (CMC) », (côté privé).

Il existe deux catégories d'indices : les premiers se réfèrent à des grandeurs simples et les secondes correspondent à des grandeurs complexes.

II- Indices simples :

1- Définition :

Les grandeurs étudiées (économiques, sociales, culturelles...) prennent des valeurs variables au cours du temps ou dans l'espace. En comparant ces différentes valeurs en terme de rapport, on obtient un indice.simple ou élémentaire.

Exemple :

Soit P_0 le prix d'un certain livre en une date t_0 et soit P_1 le prix du même livre en une deuxième date t_1.

Soit P_0=**50** DH et P_1= **60** DH.

Pour évaluer la variation du prix du livre, on peut dire qu'elle est de :

- $P_1 - P_0$= **60 −50=10** DH, soit une augmentation de $\frac{10}{50}\times 100 = 20\%$.
- $\frac{60}{50} = 1,20$ c'est-à-dire que le prix a été multiplié par $1,2$.
- $(1,20)\times 100 = 120 \Rightarrow$ valeur **100** en t_0

  $\Rightarrow$ valeur **120** en t_1

$\left(\frac{60}{50}\times 100\right)$ représente un indice de prix.

Lorsque l'indice porte sur une seule grandeur, on l'appelle indice simple ou élémentaire, il se note d'une manière générale :

Indice des prix :

$I_{t/0}=\frac{P_t}{P_0}\times 100$ avec P_t est le prix en une date t

et P_0 le prix de référence.

Indice de valeur :

$I_{t/0}=\frac{V_t}{V_0}\times 100$ avec V_t est la valeur en une date t

et V_0 la valeur de référence.

Ces indices sont exprimés en pourcentage et n'ont pas de dimension et servent par conséquent à comparer des valeurs différentes, qui ont des unités différentes. Ils ne sont pas toujours liés au temps mais peuvent être par rapport à l'espace (La ville, le pays, le lieu…) En utilisant le temps, la période de référence (0) s'appelle la période de base.

Exemple :

Supposons que le prix d'achat d'un litre de lait ait été **1,20** DH en **1965** et **2,00**DH en **1985**.

En prenant comme année de base **1965**, on a :

$$I_{1965/1985}=\frac{2}{1,20}\times 100=166\%$$

Ce qui signifie que le prix du lait a été en **1985**, **166%** celui de **1965**, ou a augmenté de **66%** de **1965** à **1985**.

En prenant comme année de référence **1985**, on a:

$$I_{1985/1965} = \frac{1,20}{2} \times 100 = 60\%$$

Ce qui signifie que le prix du lait a été en **1965**, **60%** fois celui de **1985**, ou a diminué de **40%** entre **1985** et **1965**.

Remarque :

Pour alléger l'écriture aussi bien que les quelques démonstrations qui vont suivre, nous noterons l'indice sous la forme :

$$\boxed{I_{t/0} = \frac{P_t}{P_0}}$$

en omettant le facteur multiplicatif **100**

2- Propriétés des indices élémentaires :

Les indices sont souvent l'objet de modifications ; ces modifications sont possibles grâce aux propriétés que vérifient ces indices.

a- Identité :

L'indice correspondant à une même période donnée par rapport à la même période est de **100%**.

$$I_{0/0} = I_0 = 100\%$$

b- Circularité :

Les indices élémentaires peuvent changer d'année de base par la règle de circularité :

De l'égalité $\frac{P_t}{P_0} = \frac{P_t}{P_s} \cdot \frac{P_s}{P_0}$

On déduit $$I_{t/0} = I_{t/s} \cdot I_{s/0}$$

Cette propriété permet de comparer, non seulement les périodes **0** et **t** d'une part et **0** et **s** d'autre part, mais aussi **t** et **s.**

$$I_{t/s} = \frac{I_{t/0}}{I_{s/0}}$$

La comparaison se fait indépendamment du choix de la période de référence **0.**

Exemple :

On a noté les informations suivantes concernant le prix d'un paquet de café moulu 250g, marque « X » dans une ville « Y » :

Epoques	Prix paquet de 250g (en DH)	Indices Année de base 1998
1998	11	100
1999	12,65	115
2000	15,18	138

Supposons que l'on veuille changer de base, en prenant (1999) pour période de base.

- Par application de la définition, on a :

$$I_{2000/1999} = \frac{15,18}{12,65} \times 100 = 120$$

- Par application de la propriété de circularité, on a :

$$I_{2000/1998} = I_{2000/1999} \times I_{1999/1998}$$

alors $$I_{2000/1999} = \frac{I_{2000/1998}}{I_{1999/1998}}$$

et on le multiplie par 100 pour l'exprimer en pourcentage, on obtient :

$$I_{2000/1999} = \frac{I_{2000/1998}}{I_{1999/1998}} \times 100 = \frac{138}{115} \times 100 = 120$$

c- **Réversibilité :**

On a : $\frac{P_t}{P_0} = \frac{1}{P_0/P_t}$. c'est-à-dire $\boxed{I_{t/0} = \frac{1}{I_{0/t}}}$

Ou bien $\boxed{I_{t/0} \cdot I_{0/t} = 1}$

Cette propriété est intéressante surtout lorsque l'on se réfère à un critère qui n'est plus le temps. Par exemple, si on compare le prix d'un produit au Maroc au prix du même produit en Espagne, on aura : $I_{Maroc\ /Espagne} = \frac{1}{I_{Espagne\ /Maroc}}$

d- **Transférabilité :**

La 2ème propriété entraînée par la circularité de l'indice simple est la transférabilité de l'indice. On aura :

$$\boxed{I_{t/0} = I_{t/t-1} \cdot I_{t-1/t-2} \cdot I_{t-2/t-3} \ldots I_{1/0}}$$

Ceci signifie que l'on obtient l'indice de la date **t** par rapport à **0** en faisant le produit des indices intermédiaires d'une date par rapport à l'autre.

III- Indices composés :

1- Définition :

Les grandeurs que l'on étudie, le plus souvent, sont des grandeurs composées d'un grand nombre d'éléments mesurés par des grandeurs simples. Alors, l'indice qui synthétise plusieurs indices simples s'appelle indice composé ou synthétique.

2- Emploi des moyennes :

Comme il faut aboutir à un indice unique résumant une grande quantité d'informations, on peut faire appel au calcul des moyennes pour construire ces indices.

Deux possibilités se présentent : soit on considère l'indice des moyennes, soit la moyenne des indices.

a- Indice des moyennes arithmétiques pondérées :

On note par $P_t(i)$ le prix d'un certain produit « i » à la date t et on note $Q(i)$ la quantité consommée de ce produit « i ».

Supposons que l'on veut calculer l'indice synthétique des prix de « k » produit (i=1,2,...,k) à la date t, en prenant la date 0 comme base.

Alors l'indice des moyennes arithmétiques pondérées est

$$I_{t/0}=\frac{\sum_{i=1}^{k} Q(i)\,P_t(i) \Big/ \sum_{i=1}^{k} Q(i)}{\sum_{i=1}^{k} Q(i)\,P_0(i) \Big/ \sum_{i=1}^{k} Q(i)}\times 100$$

où on a considéré les quantités $Q(i)$ comme coefficients de pondérations.

Les pondérations sont nécessaires puisque les produits composant l'indice ne doivent pas être affectés de la même mesure (par exemple, le pain dans l'indice général des prix à un poids plus important que la mayonnaise !).

b- Moyenne arithmétique pondérée des indices simples :

Dans ce cas, on détermine d'abord l'indice simple propre pour chaque produit « i », et ensuite, on fait la moyenne arithmétique pondérée des indices simples affectés à leur pondération.

Alors la moyenne arithmétique pondérée des indices simples est :

$$I_{t/0} = \frac{\sum_{i=1}^{k} Q(i) \frac{P_t(i)}{P_0(i)}}{\sum_{i=1}^{k} Q(i)} \times 100$$

Exemple :

Soient $P_t(i)$ les prix de quatre produits $i=1,2,3,4$ à une date t et soient $Q(i)$; $i=1,2,3,4$ leurs quantités consommées.

Produits	Prix unitaires (en DH)		Quantités
	$P_0(i)$	$P_t(i)$	$Q(i)$
1	10	14	4
2	12	13	5
3	15	15	3
4	3	5	8

- Calculons l'indice des moyennes arithmétiques pondérées (date de référence étant 0) :

$$I_{t/0} = \frac{(4\times14)+(5\times13)+(3\times15)+(8\times5)/4+5+3+8}{(4\times10)+(5\times12)+(3\times15)+(8\times3)/4+5+3+8}\times100$$

$$I_{t/0} = \frac{56+65+45+40}{40+60+45+24}\times100 = \frac{206}{169}\times100 = 121,9$$

Le résultat signifie qu'on dépense **121,90 DH** en *t* pour se procurer les mêmes quantités de produits qui coûtaient **100 DH** en date de base ; c'est-à-dire que l'augmentation entre ces deux dates a été de **21,9%.**

- Calculons la moyenne arithmétique pondérée des indices simples (date de base étant **0**) :

$$I_{t/0} = \frac{4\times\frac{14}{10}+5\times\frac{13}{12}+3\times\frac{15}{15}+8\times\frac{5}{3}}{4+5+3+8}\times100 = \frac{5,6+5,42+3+13,33}{20}\times100$$

donc

$$I_{t/0} = \frac{27,35}{20}\times100 = 136,75$$

Remarque :

Dans les deux indices précédents, on a supposé que les quantités consommées soient les mêmes en date *t* qu'en date de base ; alors que les quantités consommées peuvent-elles aussi varier entre ces deux dates.

1- Indices de Laspeyres (1834 - 1913) :

Pour Laspeyres, les pondérations s'effectuent pour la période de base. En distinguant les deux cas précédents, on obtient les indices suivants :

- Cas de l'indice des moyennes arithmétiques pondérées :

$$L_{t/0} = \frac{\sum_{i=1}^{k} Q_0(i)\, P_t(i)}{\sum_{i=1}^{k} Q_0(i)\, P_0(i)} \times 100$$

- Cas des moyennes arithmétiques pondérées des indices simples :

$$L_{t/0} = \frac{\sum_{i=1}^{k} \frac{P_t(i)}{P_0(i)} Q_0(i)}{\sum_{i=1}^{k} Q_0(i)} \times 100$$

- Si on considère l'indice des prix, alors c'est le rapport des dépenses totales où seuls les prix sont variables.
- Si on considère l'indice des quantités, alors c'est le rapport des dépenses totales où seules les quantités sont variables.

2- Indices de Paasche (1851 - 1925) :

Dans ce cas, les pondérations se font pour la période considérée t. Deux cas se distinguent :

- Cas de l'indice des moyennes arithmétiques pondérées :

$$P_{t/0}=\frac{\sum_{i=1}^{k}Q_t(i)P_t(i)}{\sum_{i=1}^{k}Q_t(i)P_0(i)}\times 100$$

- Cas des moyennes arithmétiques pondérées des indices simples :

$$P_{t/0}=\frac{\sum_{i=1}^{k}\frac{P_t(i)}{P_0(i)}Q_t(i)}{\sum_{i=1}^{k}Q_t(i)}\times 100$$

Tableau récapitulatif :

Indices	Laspeyres *L*	Paasche *P*
Prix *P*	$L_P=\frac{\sum Q_0 P_t}{\sum Q_0 P_0}\times 100$	$P_P=\frac{\sum Q_t P_t}{\sum Q_t P_0}\times 100$
Quantités *Q*	$L_Q=\frac{\sum P_0 Q_t}{\sum P_0 Q_0}\times 100$	$P_Q=\frac{\sum P_t Q_t}{\sum P_t Q_0}\times 100$
Valeurs globales *Vg*	$L_{Vg}=P_{Vg}=\frac{\sum Q_t P_t}{\sum Q_0 P_0}\times 100$	

Exemple :

Quatre produits ***j=1, 2, 3, 4*** sont donnés avec prix P_{ij} et quantités Q_{ij} en des dates ***i=0, t***

Produits *j*	Date 0 de base		Date *t*	
	P_{0j}	Q_{0j}	P_{tj}	Q_{tj}
1	12	6	15	7
2	5	13	8	11

3	15	9	13	18
4	8	10	10	9

Calculer les indices de Laspeyres et Paasche correspondants aux prix, aux quantités et aux valeurs globales :

Produits	P_0Q_0	P_tQ_t	P_0Q_t	P_tQ_0
1	72	105	84	90
2	65	88	55	104
3	135	234	270	117
4	80	90	72	100
Totale $\sum$	352	517	481	411

$$L_p=\frac{\sum Q_0 P_t}{\sum Q_0 P_0}\times 100=\frac{411}{352}\times 100\approx 116,76$$

$$L_Q=\frac{\sum P_0 Q_t}{\sum P_0 Q_0}\times 100=\frac{481}{352}\times 100\approx 136,65$$

$$L_{Vg}=\frac{\sum Q_t P_t}{\sum Q_0 P_0}\times 100=\frac{517}{352}\times 100\approx 146,86$$

$$P_p=\frac{\sum Q_t P_t}{\sum Q_t P_0}\times 100=\frac{517}{481}\times 100\approx 107,48$$

$$P_Q=\frac{\sum P_t Q_t}{\sum P_t Q_0}\times 100=\frac{517}{411}\times 100\approx 125,79$$

$$P_{Vg}=\frac{\sum P_t Q_t}{\sum P_0 Q_0}\times 100=\frac{517}{352}\times 100\approx 146,86$$

Remarque :

$$\frac{L_p P_Q}{100}=\frac{L_Q P_P}{100}=L_{Vg}=P_{Vg}$$

Pour l'exemple précédent,

$$\frac{(116,76)(125,79)}{100}=\frac{(136,65)(107,48)}{100}=146,87$$

3- Indices de Fisher :

Le fait de considérer la même quantité consommée pour les deux dates n'est pas satisfaisant. On peut remédier partiellement à cet inconvénient en considérant un autre indice qui est l'indice de Fischer. Il est donné par la moyenne géométrique des indices de Laspeyres et de Paasche :

$$F_p=\sqrt{L_p P_p} \qquad \text{(Prix)}$$

$$F_Q=\sqrt{L_Q P_Q} \qquad \text{(Quantités)}$$

4- Propriétés des indices composés :

- L'indice des moyennes arithmétiques pondérées des différentes époques est réversible et transférable (à la condition que la pondération ne change pas d'une période à l'autre).
- L'indice composé, moyenne arithmétique pondérée des indices simples n'est généralement ni réversible, ni transférable.
- Aucun des trois indices (de Laspeyres, de Paasche, de Fisher) ne possède la propriété de circularité.
- L'indice de Fisher est réversible.

IV- Les indices dans la vie économique :

Il est de plus en plus nécessaire, pour gérer les affaires d'un pays, de connaître, à tout moment, la situation économique de celui-ci.
Toute évolution doit être suivie avec régularité d'où la nécessité de calculer des indices analogues à ceux que nous avons introduits dans ce chapitre.

Au Maroc, La Direction de la Statistique mène des enquêtes périodiques dans le domaine économique et social. Les données obtenues à partir de ces enquêtes servent à l'élaboration des indices qui permettent de suivre la conjoncture économique du pays. La formule employée pour le calcul des indices au Maroc est celle de **Laspeyres**.

Une étude approfondie de ces indices, publiés par la Direction de la statistique, nous entraînerait dans des détails que nous ne pouvons pas développer ici sans alourdir considérablement ce chapitre. Mais ceci montre l'importance prise par les **indices synthétiques** dans la vie économique.

Aucune prévision, aucun plan de développement ne saurait se faire sans les prendre en considération.

Exercices

Exercice 1 :

On dispose des informations sur les salaires moyens d'une entreprise industrielle secteur par secteur et son nombre de salariés :

Secteurs	Salaire moyen en t_0	Salaire moyen en t_1	Nombre en t_0	Nombre en t_1
1	5800	6000	250	241
2	6200	6250	300	320
3	5920	6080	120	122
4	6150	6300	45	43

L'entreprise désire connaître l'évolution générale de ses salaires entre t_0 et t_1.

Calculer les indices de Laspeyres et de Paasche.

Exercice 2 :

On donne les prix et les quantités de deux produits A et B dans deux villes 0 et 1.

Produits	Ville 0		Ville 1	
	Prix	Quantités	Prix	Quantités
A	100	100	10	80
B	10	50	200	80

1) Calculer l'indice de Laspeyres des quantités de la ville 1 par rapport à la ville 0. D'après cet indice, quelle est la ville qui a le plus produit ?

2) Calculer un nouvel indice de Laspeyres, en prenant pour base la ville 1. Peut-on conclure de la même manière qu'au 1) ?

3) Calculer les indices de Paasche de bases les villes 0 et 1. Que peut-on conclure ?

Exercice 3 :

On donne la moyenne des prix de gros et la production de lait, de beurre et de fromage dans un pays donné pour les années 1999, 2000 et 2001.

Produits	Prix (par Kilo ou litre)			Quantités (millions de Kilo)		
	1999	2000	2001	1999	2000	2001
Lait	0,395	0,389	0,413	9675	9717	10436
Beurre	6,15	6,22	5,97	117,7	115,5	115,5
Fromage	3,48	3,54	3,89	77,93	74,39	82,79

1) Calculer un indice global des prix de gros de ces produits pour l'année 2001 :
   - En choisissant 1999 comme année de base.
   - En choisissant 1999-2000 comme période de base.
2) Calculer l'indice de Laspeyres des prix pour l'année 2001 :
   - En choisissant 1999 comme année de base.
   - En choisissant 1999-2000 comme période de base.
3) Calculer l'indice de Paasche des prix pour l'année 2001 :
   - En choisissant 1999 comme année de base.

   En choisissant 1999-2000 comme période de base.

Annexe :

Interpolation linéaire

Effectuer une interpolation linéaire entre deux points *A(a, f(a))* et *B(b, f(b))* de la courbe représentative C_f d'une fonction *f*, c'est supposer qu'entre ces deux points la fonction peut être remplacée par une fonction linéaire (qu'il faudra déterminer), afin d'effectuer d'éventuelles prévisions des valeurs de *f* inconnues.

En statistique, souvent, on a des données qui manquent ou aberrantes. Alors, l'interpolation linéaire nous permet de retrouver une valeur approchée de telles valeurs.

Par exemple, soit la situation suivante :

$[e_{i-1}, e_i[$	n_i
⋮	⋮
[3000, 3500[	15
[3500, 4000[	20
⋮	⋮

On veut avoir la valeur ni qui correspond à la classe [3000, 3200[.

Alors, on a, par interpolation linéaire $\frac{x-15}{20-15} = \frac{3200-3000}{3500-3000}$

$$\Rightarrow x = \frac{200}{500} \times 5 + 15$$

$$\Rightarrow x = 17$$

Bibliographie :

- A. Dadoun : « Précis de mathématiques générales, cours et exercices résolus, Tome 2 », Bréal, 1994.
- B. Py, « Exercices corrigés de Statistique Descriptive », Economica, 2ème édition, 1994.
- B. Grais: « Statistique descriptive: techniques statistiques.1 », Dunod, 3ème édition, 1994.
- C. Dhuin: « Problèmes corrigés de statistiques posés aux examens du DEUG de sciences économiques (1° et 2° années) », Ellipses, 1991.
- G. Calot, « Cours de Statistique Descriptive », Dunod, 1973.
- G. Chauvat & J-Ph. Rénau: « Statistiques descriptive : exercices corrigés » , Armand Colin, 1996.
- J. L. Bon & P. Leterrier : « Mathématiques », Vuibert, 1989.
- P. Degrelle & A. Vandeville : « Méthodes quantitatives », Techniplus, 1993.
- R. Barthe : « La statistique descriptive en 10 Leçons : Méthode progressive ABCD », Economica, 1989.
- R. Giraud : « Méthodes quantitatives : Mathématique-Statistique-Probabilités, cours et exercices » Vuibert, 1989.
- W. Maséri : « Statistique et Calcul des probabilités», Editions Sirey, Dalloz, 1996.

Cet ouvrage de Statistique Descriptive s'adresse aux étudiants de la première année universitaire, premier et deuxième semestres de toutes les filières de la nouvelle génération des Licences Fondamentales en sciences, sciences économiques et gestion...ainsi qu'aux étudiants de la première année des écoles d'ingénieurs et des écoles supérieures de commerce et gestion.

C'est un support de cours et des exercices, présentant les notions de base de la Statistique. Il dispense, donc, les étudiants de prendre des notes de cours, néanmoins, il les invite à faire un bon nombre d'exercices, extraits des contrôles et des examens, présentés à la fin de chaque chapitre.

Mohamed El Merouani est docteur en Sciences Mathématiques, Enseignant chercheur et chef de département de Statistique et Informatique à la Faculté Polydisciplinaire de Tétouan